Online Learning for Decentralized Multi-Agent Planning in Repeated Hedonic Skill Games

Jaber Valizadeh[1], **Ray Telikani** [2]

[1]School of Computing, Data and Mathematical Sciences, Western Sydney University, Sydney, NSW, Australia
[2] Data Science Institute, University of Technology Sydney, Sydney, NSW, Australia
j.valizadeh@westernsydney.edu.au, ray.telikani@uts.edu.au

Abstract

Coalition formation is a fundamental capability in decentralized multi-agent planning, where heterogeneous agents must coordinate to execute tasks whose feasibility depends on complementary capabilities. *Hedonic Skill Games* offer a compact model for representing such structured interactions, but existing results rely on the unrealistic assumption that agents possess full knowledge of both task requirements and the skills of their coalition members. This assumption breaks down in many planning domains, such as crowdsourced task allocation or distributed multi-robot task execution, where agents must plan under uncertainty and learn from partial observations. We introduce *repeated* Hedonic Skill Games in which agents repeatedly form coalitions, execute feasible tasks, and receive only bandit feedback corresponding to their realized utilities. We develop an Upper Confidence Bound (UCB)-driven online learning algorithm that enables decentralized agents to jointly plan coalition choices despite incomplete information, balancing exploration of unknown coalitions with exploitation of realized utilities. We show that the resulting dynamics achieve sublinear *Nash regret* and converge to ε-approximate Nash equilibria. Experiments on synthetic and real-world problem instances demonstrate convergence behavior, improved social welfare, and the practicality of the approach for large-scale distributed planning scenarios.

Code: github.com/HedonicSkillGame.git
Datasets: autonomousrobots.nl/paper_websites/sadcher_MRTA/

Introduction

Coordinating autonomous agents to execute complex tasks in decentralized settings is a fundamental challenge in multi-agent planning (Von Martial 1992). Real-world domains such as human–robot teamwork, disaster response, and crowdsourcing platforms require agents with diverse and complementary capabilities to self-organize into coalitions capable of jointly accomplishing tasks that no single agent can perform alone (Valizadeh, Zhang, and Mubin 2024). Achieving efficient, stable coordination without centralized control remains an open problem at the intersection of automated planning and multi-agent systems.

Hedonic games offer a well-established framework for coalition formation, in which each agent's utility depends solely on the coalition they are part of (Drèze and Greenberg 1980), without caring about the structure of other coalitions, i.e., externalities are ignored. However, classical hedonic games are ill-suited for planning domains: they ignore task feasibility constraints, skill requirements, and the structure of the planning problem itself. Hedonic Skill Games (HSGs) (Gourvès and Monaco 2024) close this critical gap by explicitly modeling agent skills and task requirements. In an HSG, a coalition generates value only if its aggregated skills satisfy the demands of one or more tasks, and individual utilities are derived from the feasible task allocations enabled by the coalition's composition.

Existing work on HSGs has primarily analyzed structural properties, the existence and computation of stable outcomes (e.g., core or Nash-stable partitions), and complexity under complete information (Gourvès and Monaco 2024; Gourvès and Monaco 2025). These assumptions, however, do not reflect the reality of decentralized planning environments, where agents lack global knowledge of others' skills, task requirements, or coalition payoffs. Instead, agents must learn from experience: they form coalitions, attempt task execution, and observe only their own realized utilities by receiving bandit feedback from an online, repeated setting.

We propose a new framework for *online learning* in coalition formation from a non-cooperative viewpoint, where coalition outcomes are assessed using *Nash equilibrium*. To address uncertainty in decentralized planning, we introduce the *repeated hedonic skill game (RHSG)*, a learning extension of HSGs in which agents repeatedly choose coalitions, execute feasible tasks, and learn solely from their own observed utilities. The central objective is to develop a learning procedure that guides these agents toward high-welfare, stable coalition structures despite limited information and the absence of coordination.

Our main contribution is `UCB-NE`, an optimistic online learning algorithm based on the Upper Confidence Bound (UCB) principle. Because agents in RHSG lack knowledge of the skills of others, and task requirements (and observe only bandit feedback), UCB provides a principled way to manage this uncertainty. Each agent maintains a confidence interval over the expected utility of every coalition it could join and selects the coalition whose upper confidence bound is maximal. This *optimism-in-the-face-of-uncertainty* approach provides a principled mechanism for balancing ex-

ploration with exploitation. Crucially, it enables agents to independently and efficiently learn the value of coalition choices without any communication or centralized coordination.

We prove that in key subclasses of repeated HSGs, UCB-NE attains provably sublinear Nash regret (Ding et al. 2022; Liu et al. 2021a) and converges with high probability to ε-approximate Nash equilibria, despite agents receiving only bandit feedback. These results give the first provably efficient decentralized learning method for forming coalitions when agents do not know others' skills or the tasks in advance. Our experiments on synthetic and real-world-inspired problems support the theory: UCB-NE finds stable, high-welfare coalitions and clearly outperforms standard baselines.

Related Work

The game studied in this study falls into the family of utility-based hedonic games, where agents' preferences over coalitions are represented numerically via utility functions, allowing for quantitative comparisons of outcomes.

Several variants of utility-based hedonic games have been introduced to capture different preference structures among agents, including *Additively Separable Hedonic Games* (Bogomolnaia and Jackson 2002), *Fractional Hedonic Games* (Aziz et al. 2019), and *Hedonic Project Games* (Valizadeh, Zhang, and Mubin 2025b). Extensive work has examined expressive preference representations (Aziz and Savani 2016), existence of stable outcomes (Aziz and Brandl 2012; Bilò et al. 2018), and algorithmic aspects (Brandt, Bullinger, and Wilczynski 2021).

Hedonic Skill Games (Gourvès and Monaco 2024) analyze how self-interested agents form coalitions based on complementary abilities, without transferring utility between members. In the aforementioned game models, it is typically assumed that agents' preferences are completely known, enabling the identification of stable coalition structures. However, this assumption presents a significant limitation when applying such models to real-world scenarios, where agents often operate with incomplete or uncertain knowledge of others' preferences. To handle this issue, *learning-based approaches*, such as *PAC-stability* (Sliwinski and Zick 2017) and *online learning* (Cohen and Agmon 2024, 2025), have been introduced. In online learning, agents must be partitioned at each time step to minimize cumulative regret.

Bandit algorithms and reinforcement learning techniques have been applied to decentralized matching and coalition formation problems (Liu et al. 2021a,b), providing regret bounds under various feedback settings. Online learning hedonic coalition formation games builds upon these foundations by integrating them with dynamic coalition formation (Cohen and Agmon 2024, 2025). It also draws from general online learning theory (Cesa-Bianchi and Lugosi 2006) and regret minimization in potential games (Ding et al. 2022), enabling convergence guarantees under bandit feedback with unknown agent preferences.

Preliminaries

Given any integer q, we denote by $[q]$ the set $\{1, \dots, q\}$.

We consider *hedonic skill games* (Gourvès and Monaco 2024) (see also (Bachrach, Markakis, and Zuckerman 2013) and (Bachrach and Rosenschein 2008)), that involves a finite set of agents $N = \{1, 2, \dots, n\}$, a finite set of skills $\mathcal{S} = \{1, 2, \dots, k\}$, and a finite set of tasks $M = \{j_1, \dots, j_m\}$. Let $S_i \subseteq \mathcal{S}$ be a non-empty skill set for each agent $i \in N$, and $S_j \subseteq \mathcal{S}$ be a non-empty set of skills required for each task $j \in M$. Furthermore, each task $j \in M$ is associated with a positive value determined by a value function $v : M \to \mathbb{R}_+$, where $v(j)$ denotes the value of task j.

A *coalition* is defined as any non-empty subset $C \subseteq N$, and a *coalition structure* (or partition) is a set of such disjoint, non-empty coalitions that together cover the entire agent set N. For any $\ell \in [q]$, we overload $S(C_\ell) = \bigcup_{i \in C_\ell} S_i$ as the aggregate skill set of a ℓ'th coalition. A coalition is capable of executing a task j if the required skill set S_j is contained within $S(C_\ell)$. Accordingly, we define $\mathbf{M}(C_\ell) = \{j \in M \mid S_j \subseteq S(C_\ell)\}$ as the set of tasks that coalition C_ℓ is capable of performing. Each coalition executes all of its tasks exactly once. The value $v(j)$ derived from executing a task j is distributed exclusively among those agents in the coalition who contribute at least one of the required skills for j.

Each agent $i \in N$ selects a coalition by choosing a *strategy* $x_i \in [q]$, where $q \leq n$ represents the maximum number of possible coalitions. A *strategy profile* is then denoted by $\mathbf{x} = (x_1, \dots, x_n) \in [q]^n$. Every strategy profile $\mathbf{x}$, induces a coalition structure $\pi(\mathbf{x}) = \{C_\ell(\mathbf{x}) \mid \ell \in [q], C_\ell(\mathbf{x}) \neq \emptyset\}$, where each coalition formed under $\mathbf{x}$ is defined as $C_\ell(\mathbf{x}) = \{i \in N \mid x_i = \ell\}$. The induced coalition structure satisfies both full coverage, $\bigcup_{\ell \in [q]: C_\ell(\mathbf{x}) \neq \emptyset} C_\ell(\mathbf{x}) = N$, and mutual exclusivity, $C_h(\mathbf{x}) \cap C_g(\mathbf{x}) = \emptyset$ for all $h, g \in [q]$ with $h \neq g$. Given a strategy profile $\mathbf{x}$ and related coalition structure $\pi(\mathbf{x})$, the specific coalition to which agent i belongs is denoted by $\pi_i(\mathbf{x})$.

The value of task j is divisible and equally shared among its required skills ($|S_j|$), i.e., each skill receives an amount of $v(j)/|S_j|$. For each skill $s \in S_j$, this share is then split equally among the agents in the coalition who possess that skill. Thus, if coalition $C_\ell(\mathbf{x})$ performs task j, only agents with at least one required skill in S_j receive a share of its value.

Formally, given a strategy profile $\mathbf{x} \in [q]^n$, for each $s \in S_j$, every agent $i \in C_\ell(\mathbf{x})$ with $s \in S_i$ obtains a reward of

$$\frac{v(j)}{|S_j| \cdot |\{i' \in C_\ell(\mathbf{x}) \mid s \in S_{i'}\}|}$$

where $|S_j| \geq 1$ is the number of required skills for executing task j, and $|\{i \in C_\ell(\mathbf{x}) \mid s \in S_i\}|$ represents the number of agents within the coalition $C_\ell(\mathbf{x})$ who possess the skill $s \in S_j$. The total sum of the rewards received by an agent constitutes her utility. The utility of an agent $i \in N$ under strategy profile $\mathbf{x} \in [q]^n$ is given by:

$$u_i(\mathbf{x}) = \sum_{j \in \mathbf{M}(\pi_i(\mathbf{x}))} \sum_{s \in S_j \cap S_i} \frac{v(j)}{|S_j| \cdot n_s(\pi_i(\mathbf{x}))} \qquad (1)$$

where $n_s(\pi_i(\mathbf{x})) = |\{i' \in \pi_i(\mathbf{x}) \mid s \in S_{i'}\}| \geq 1$ is the number of agents in coalition $\pi_i(\mathbf{x})$ that possess the same skill s as i (and $n_s(\pi_i(\mathbf{x})) \geq 1$ whenever j is executable by the coalition $\pi_i(\mathbf{x})$). If $\mathbf{M}(\pi_i(\mathbf{x})) = \emptyset$, then $u_i(\mathbf{x}) = 0$.

The following example illustrates how agents, skills, and tasks interact in a hedonic skill game, and how utilities are computed under different strategy profiles.

Example 1. Consider a game with skill set $\mathcal{S} = \{a, b, c\}$ and two tasks $M = \{j_1, j_2\}$, and four agents $N = \{1, 2, 3, 4\}$ are reported in Tables 1 and 2.

Task	Required Skills	Value
j_1	$\{a, b\}$	12
j_2	$\{b, c\}$	18

Table 1: Tasks, required skills, and values.

Agent	Skill Set
1	$\{a\}$
2	$\{a, b\}$
3	$\{b\}$
4	$\{c\}$

Table 2: Agents and their skills.

We compare two strategy profiles:

$$\mathbf{x} = (1, 1, 1, 2) \qquad \text{and} \qquad \mathbf{x}' = (1, 2, 2, 2),$$

which induce distinct coalition structures, executable tasks, and utilities.

Profile $\mathbf{x} = (1, 1, 1, 2)$. The induced coalitions are

$$C_1(\mathbf{x}) = \{1, 2, 3\}, \qquad C_2(\mathbf{x}) = \{4\}.$$

Coalition $C_1(\mathbf{x})$ has skills $\{a, b\}$ and can execute task j_1, while $C_2(\mathbf{x})$ cannot execute any task. For task j_1, each required skill receives $12/2 = 6$. Agents possessing skill a (agents 1 and 2) each obtain $6/2 = 3$, and those possessing skill b (agents 2 and 3) each obtain $6/2 = 3$. Thus,

$$u_1(\mathbf{x}) = 3,\ u_2(\mathbf{x}) = 6,\ u_3(\mathbf{x}) = 3,\ u_4(\mathbf{x}) = 0.$$

Profile $\mathbf{x}' = (1, 2, 2, 2)$. The induced coalitions are

$$C_1(\mathbf{x}') = \{1\}, \qquad C_2(\mathbf{x}') = \{2, 3, 4\}.$$

Coalition $C_1(\mathbf{x}')$ performs no task, while $C_2(\mathbf{x}')$ can execute both j_1 and j_2.

For j_1 (value 12), each skill receives $12/2 = 6$. Skill a is provided solely by agent 2 (reward 6), while skill b is provided by agents 2 and 3 (reward 3 each).

For j_2 (value 18), each skill receives $18/2 = 9$. Skill b is provided by agents 2 and 3 (reward 4.5 each), and skill c is provided by agent 4 (reward 9).

Summing contributions yields:

$$u_1(\mathbf{x}') = 0,\ u_2(\mathbf{x}') = 13.5,\ u_3(\mathbf{x}') = 7.5,\ u_4(\mathbf{x}') = 9.$$

This example illustrates how agents' strategic choices shape coalition structures, determine which tasks can be executed, and ultimately affect the distribution of utilities.

In many real-world settings, agents often operate without prior knowledge of task requirements or the capabilities of other participants. Instead, this information is gradually learned through repeated interactions and feedback. Motivated by these practical constraints, we extend our model to a repeated game framework in which agents begin with incomplete information.

Repeated Hedonic Skill Games with Bandit Feedback

We now extend hedonic skill games to an online, repeated setting in which agents have no prior knowledge of other agents' skills or task requirements. Instead, agents learn through repeated interactions with stochastic feedback. We assume that agents' utilities are within $[0, u_{\max}]$, where $u_{\max} > 0$ is a known upper bound on the maximum possible utility for any agent. The game proceeds over discrete rounds $t = 1, 2, \ldots, T$. Every strategy profile $\mathbf{x}$ sampled at time t, induces a coalition structure $\pi(\mathbf{x})$ exactly as in the static game. Agent i observes only her realized utility $u_i^t(\mathbf{x}) \in [0, u_{\max}]$ (bandit feedback), and whose mean is $\bar{u}_i(\mathbf{x})$; no other information is revealed. The agents' goal is to learn, through this repeated bandit feedback, coalition choices that lead to high-welfare and stable outcomes.

Agents may employ mixed strategies for exploration: at round t, agent i samples $x_i^t \sim \sigma_i^t \in \Delta([q])$, where $\Delta([q])$ is the probability simplex over $[q]$. The expected utility of agent i under a mixed strategy profile $\sigma^t = (\sigma_1^t, \ldots, \sigma_n^t)$ is

$$U_i(\sigma^t) = \mathbb{E}_{\mathbf{x} \sim \sigma^t}\left[\bar{u}_i(\mathbf{x})\right] \tag{2}$$

where the expectation is over the product distribution induced by the individual mixed strategies.

Stability Concepts The existence of Nash equilibria is crucial for ensuring stability in the game, as it implies that no player has an incentive to unilaterally deviate from their strategy, given others' strategies (Valizadeh, Zhang, and Mubin 2025a). A pure strategy profile $\mathbf{x} \in [q]^n$ is a *Nash equilibrium* (NE) of the underlying complete-information HSG, if no agent can strictly increase her utility by unilaterally changing her coalition.

Definition 1. *Given a game instance $\mathcal{G}$, a strategy profile $\mathbf{x} = (x_i, \mathbf{x}_{-i}) \in [q]^n$ is a pure Nash equilibrium if, for each player $i \in N$, and for any alternative $\ell \in [q]$,*

$$u_i(\mathbf{x}) \geq u_i(\ell, \mathbf{x}_{-i})$$

Because agents play mixed strategies and only observe expected utilities, we focus on convergence to approximate Nash equilibria of the underlying game. At each time t, for each player i, $\sigma_i^{\star,t} \in \Delta([q])$ best response to the other players' mixed strategies satisfying $U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) = \max_{\sigma_i \in \Delta([q])} U_i(\sigma_i, \sigma_{-i}^t)$. Thus, for any $\varepsilon \geq 0$, the mixed strategy profile σ^t sampled at time t is ε-approximately Nash equilibrium (ε-NE) if

$$\max_{i \in N} \left(U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t)\right) \leq \varepsilon$$

Regret Minimization To evaluate the learning performance, we adopt the standard notion of *regret* (Valizadeh, Zhang, and Mubin 2025b; Lattimore and Szepesvári 2020). Regret quantifies the average difference between the expected utility an agent could have achieved by always playing her best possible strategy (i.e., the best response to others' strategies based on current estimates) and the expected utility she actually obtained by following her adaptive strategy over all rounds of the game (Zinkevich 2007; Liu et al. 2021b).

Definition 2. *Given a sequence of mixed strategy profiles* $\mathbf{x}$*, the* average Nash regret *of agent* i *after* T *rounds is*

$$\mathcal{R}_i^T = \max_{i \in N} \frac{1}{T} \sum_{t=1}^{T} \Big(U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t) \Big)$$

We define the *total average Nash regret* over all agents as $\mathcal{R}^T = \sum_{i \in N} \mathcal{R}_i^T$. A learning algorithm is said to achieve *sublinear regret* if $\mathcal{R}_i^T = \mathcal{O}(1)$ as $T \to \infty$, implying that the per-round regret converges to zero over time (Lattimore and Szepesvári 2020).

Our main algorithm `UCB-NE` will be shown to achieve sublinear Nash regret with high probability in important subclasses of repeated HSGs, guaranteeing convergence to approximate Nash equilibria despite severe information constraints and stochastic rewards.

Previous research has shown that sublinear regret bounds are attainable even in settings where agents cannot compute exact gradients (Zinkevich 2007; Liu et al. 2021b; Ding et al. 2022; Cohen and Agmon 2025). Building on these insights, a central goal of this work is to develop a learning algorithm for repeated hedonic skill games that minimizes average Nash regret in environments characterized by limited feedback and incomplete information. In particular, we aim to achieve regret bounds that are sublinear in the time horizon T and polynomial in the size of the coalition space, which grows exponentially with the number of agents.

Optimism Principle and Feedback Settings In our setting, agents select coalitions in an *uncertain* environment where initially agents lack full knowledge of both the task requirements and the skill sets of other agents. To navigate this uncertainty, we introduce a UCB-based algorithm according to the principle of *optimism in the face of uncertainty*[1], which states that one should act as if the environment is as nice as *plausibly possible* (Lattimore and Szepesvári 2020). At each time step t, and for any strategy profile profile $\mathbf{x}$, we assume that each agent i can receive *bandit* feedback, as commonly studied in multi-agent online learning (Cui et al. 2022; Jones, Nguyen, and Nguyen 2023; Cohen and Agmon 2025).

Bandit Feedback Under bandit feedback, each agent $i \in N$ observes only her realized utility $u_i^t(\mathbf{x})$ at round t, corresponding to the sampled joint action $\mathbf{x} \sim \sigma^t$. Crucially, agent i does not observe the utilities associated with alternative coalitions she could have joined, nor the utilities or actions of other agents. At time step $t = 1$, each agent i initializes an arbitrary $\sigma_i^t \in \Delta[q]$. In each round t, agent i selects a coalition $x_i \in [q]$ that maximizes her UCB estimate, leading to a strategy profile $\mathbf{x} = (x_1, \ldots, x_n)$. Once the coalition structure $\pi(\mathbf{x})$ is formed, agent i observes the outcome $u_i^t(\mathbf{x})$ from her coalition $\pi_i^t(\mathbf{x})$.

Agent i maintains an empirical estimate of her utility for each coalition $\ell \in [q]$ by averaging the feedback she has received in past rounds. Let $f_{i,\ell}^t(\mathbf{x}) = \sum_{\tau=1}^{t} \mathbf{1}\{x_i^\tau = \ell\}$ denote the number of times that agent i selects ℓ'th coalition up to time t, where, for any time $\tau \in [t]$, $\mathbf{1}\{x_i^\tau = \ell\}$ equals 1 if $x_i^\tau = \ell$, and 0 otherwise. Then, the empirical mean utility of agent i for ℓ'th candidate coalition up to time t is

$$\hat{u}_i^t(\mathbf{x}) = \frac{\sum_{\tau=1}^{t} u_i^\tau(\mathbf{x}) \cdot \mathbf{1}\{x_i^\tau = \ell\}}{f_{i,\ell}^t(\mathbf{x}) \vee 1} \tag{3}$$

Agents use this feedback to guide future exploration and exploitation. To balance these, each agent constructs a UCB estimate that combines the empirical utility with an exploration bonus. Then the UCB estimate for agent i under profile $\mathbf{x}$ is $\hat{\mathcal{U}}_i^t(\mathbf{x}) = \hat{u}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$, where $\hat{b}_i^t(\mathbf{x})$ is the exploration bonus that is given by:

$$\hat{b}_i^t(\mathbf{x}) = \sqrt{\frac{n^3 \log(4(n^2+1)T/\delta))}{f_{i,\ell}^t(\mathbf{x}) \vee 1}} \tag{4}$$

Here, $\delta \in (0, 1]$ is a confidence parameter that controls the trade-off between exploration and exploitation (Lattimore and Szepesvári 2020) for more discussion. The bonus term decreases as more information about a task is collected, encouraging agents to explore unfamiliar tasks while favoring those with historically higher observed utility.

Regret Bounds and Convergence Guarantees

We derive regret and convergence guarantees for the `UCB-NE` algorithm in repeated hedonic skill games under bandit feedback for the general case (arbitrary skill sets $|S_i| \geq 1$ for agents $i \in N$ and arbitrary task requirements $|S_j| \geq 1$ for tasks $j \in M$). Unlike the special cases of singleton tasks or singleton agents considered in prior work (Gourvès and Monaco 2025), where a pure Nash equilibrium is always guaranteed to exist, the general hedonic skill game may not admit any pure Nash equilibrium. However, by the fundamental theorem of finite games (Nash Jr 1950), at least one mixed-strategy Nash equilibrium always exists. We therefore extend the regret analysis and convergence results from the pure-strategy setting of these special cases to the general repeated game, where agents employ mixed strategies and the natural solution concept is an (approximate) mixed Nash equilibrium. The theoretical guarantees established in Theorems 1 and 2 thus hold for the full class of repeated hedonic skill games, ensuring sublinear Nash regret and convergence to ε-approximate mixed Nash equilibria with high probability, even when pure-strategy stable outcomes may not exist.

[1]For bandits, the *optimism principle* means using the data observed so far to assign to each coalition a value, called the *upper confidence bound*, that with high probability overestimates the unknown expected utility

Algorithm 1: UCB−NE for HSGs

Input: Number of tasks m, number of agents n, number of skills k, time horizon T, confidence parameter $\delta \in (0, 1]$, task values v, and $\varepsilon \in (0, 1]$
Output: Mixed strategy profile σ^{T+1} that is an ε-NE

Initialize preference $\sigma_i^1 \in \Delta[q]$ for each agent $i \in N$
for each round $t = 1$ to T **do**
  for each agent $i \in N$ **do** ▷ Phase 1: All agents sample
    Sample strategy $x_i \sim \sigma_i^t$
  end for
  Form strategy profile $\mathbf{x} = (x_1, \ldots, x_n)$ ▷ Phase 2: Form joint profile
  Induce coalition structure $\pi(\mathbf{x})$
  for each agent $i \in N$ **do** ▷ Phase 3: Observe and update
    Observe realized utility $u_i^t(\mathbf{x})$ from coalition $\pi_i(\mathbf{x})$
    Update empirical mean $\hat{u}_i^t(\mathbf{x})$ by Eq (3)
    Compute UCB estimate $\hat{\mathcal{U}}_i^t(\mathbf{x}) = \hat{u}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$ by Eq (4)
  end for
  Compute $\sigma^{t+1} \leftarrow \varepsilon$-BRD$(\hat{\mathcal{U}}^t, \varepsilon)$ using Algorithm 2
end for

Theoretical Analysis

The theoretical analysis shows that UCB−NE can guide agents toward stable coalition choices even when they only observe limited feedback. In settings where a pure Nash equilibrium is guaranteed, the algorithm reaches an ε-approximate equilibrium within a fixed number of rounds and achieves sublinear Nash regret. This means that, over time, agents have less and less reason to change their decisions, and the system naturally moves toward stable and efficient coalition structures. The following lemma shows that, in these special cases (either singleton tasks or singleton agents), a pure Nash equilibrium is always guaranteed to exist.

Lemma 1 ((Gourvès and Monaco 2024)). *In hedonic skill games with either singleton tasks ($|S_j| = 1$, for all $j \in M$) or singleton agents ($|S_i| = 1$, for all $i \in N$), a pure Nash equilibrium always exists.*

Lemma 1 guarantees the existence of a pure Nash equilibrium in hedonic skill games when either tasks or agents possess only a single skill. However, in the general setting (where tasks may require multiple skills and agents may possess multiple skills), such stable outcomes are no longer guaranteed. In these richer and more expressive cases, pure Nash equilibria may fail to exist entirely. To address this, we turn to mixed strategies, under which an ε-approximate Nash equilibrium is always guaranteed to exist.

Indeed, every finite game admits at least one mixed Nash equilibrium (Nash Jr 1950), though computing one is known to be PPAD-complete (Daskalakis, Goldberg, and Papadimitriou 2009). Theorem 1 shows that the proposed UCB−NE algorithm achieves sublinear Nash regret in repeated hedonic skill games. Consequently, even in the absence of pure equilibria, the learning dynamics provably converge to an ε-approximate mixed Nash equilibrium, effectively balancing exploration and exploitation throughout the learning process.

Algorithm 2: ε-BRD for HSGs

Input: UCB estimates $\hat{\mathcal{U}}^t$; accuracy $\varepsilon \in (0, 1]$
Output: $\sigma^{\tau+1}$

Initialize $\sigma^1 = \mathbf{x}^1$: arbitrary deterministic join policy
for $\tau = 0, 1, 2, \ldots, \left\lceil \frac{n.u_{\max}}{\varepsilon} \right\rceil$ **do**
  for each agent $i \in N$ **do**
    $\Delta_i^\tau \leftarrow \max_{x_i \in [q]} \hat{\mathcal{U}}_i^t(x_i, \sigma_{-i}^\tau) - \hat{\mathcal{U}}_i^t(\sigma^\tau)$
    $x_i^{\tau+1} \leftarrow \arg\max_{x_i \in [q]} \hat{\mathcal{U}}_i^t(x_i, \sigma_{-i}^\tau) - \hat{\mathcal{U}}_i^t(\sigma^\tau)$
  end for
  if $\max_{i \in N} \Delta_i^\tau \le \varepsilon$ **then**
    return σ^τ
  end if
  $i^\star \leftarrow \arg\max_{i \in N} \Delta_i^\tau$
  $\sigma^{\tau+1}(i^\star) \leftarrow x_{i^\star}^{\tau+1}$; $\sigma^{\tau+1}(i) \leftarrow \sigma^\tau(i)$ for all $i \neq i^\star$
end for

Theorem 1. *In repeated hedonic skill games, for any $\varepsilon \in (0, 1]$, and for any initial mixed strategy σ^1, there exists a time step $T_0 \le \left\lceil \frac{n.u_{\max}}{\varepsilon} \right\rceil$, such that*

$$\max_{i \in N} \big(U_i(\sigma_i^{\star, T_0}, \sigma_{-i}^{T_0}) - U_i(\sigma^{T_0}) \big) \le \varepsilon$$

Proof. Let $N = \{1, \ldots, n\}$ be the set of agents, and let $U_i(\sigma)$ denote the expected utility of agent i under the mixed profile σ. Define the expected social welfare

$$\mathrm{SW}(\sigma) = \sum_{i \in N} U_i(\sigma)$$

Since each $U_i(\sigma) \in [0, u_{\max}]$, for all mixed profiles σ, we have

$$0 \;\le\; \mathrm{SW}(\sigma) \;\le\; n.u_{\max} \tag{5}$$

For each time step t, define the maximum unilateral gain

$$\Delta_t = \max_{i \in N} \Big(U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t) \Big)$$

so that the profile σ^t is an ε-Nash equilibrium if and only if $\Delta_t \le \varepsilon$.

Suppose that at some round t we have $\Delta_t > \varepsilon$. Then there exists an agent i and a (mixed) best response $\sigma_i^{\star,t}$ such that

$$U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t) > \varepsilon$$

If we modify only agent i's strategy in σ^t to obtain the profile $(\sigma_i^{\star,t}, \sigma_{-i}^t)$ then all other agents' expected utilities remain unchanged. Hence, the expected social welfare increases by more than ε:

$$\mathrm{SW}(\sigma_i^{\star,t}, \sigma_{-i}^t) - \mathrm{SW}(\sigma^t) = U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t) > \varepsilon \tag{6}$$

Consider the sequence of mixed profiles $\sigma^1, \sigma^2, \ldots$. Every time $\Delta_t > \varepsilon$, Eq (6) shows that the expected social welfare increases by at least ε. Since SW is bounded above by

$n.u_{\max}$ from Eq (5), such strict improvements can occur at most $\left\lceil \frac{n.u_{\max}}{\varepsilon} \right\rceil$ times.

Therefore, within the first $T_0 \leq \left\lceil \frac{n.u_{\max}}{\varepsilon} \right\rceil$ rounds, there must exist a time step T_0 for which $\Delta_{T_0} \leq \varepsilon$. Equivalently,

$$\max_{i \in N} \big(U_i(\sigma_i^{\star,T_0}, \sigma_{-i}^{T_0}) - U_i(\sigma^{T_0}) \big) \leq \varepsilon$$

□

To analyze the Nash regret of the UCB−NE algorithm, we first establish a high-probability concentration bound for each agent's empirical utility estimates under bandit feedback. Since agents only observe the realized utility of the coalition they join, accurate estimation of expected utilities is essential for ensuring that UCB-based exploration correctly identifies improving deviations. The following lemma shows that, with high probability, the empirical utility of every agent for every candidate coalition remains within the corresponding confidence interval defined by the UCB bonus, uniformly over all rounds.

Lemma 2. *For any $\delta \in (0, 1]$, with probability at least $1-\delta$, the following holds simultaneously for all rounds $t \leq T$, all agents $i \in N$, and all strategy profiles $\mathbf{x} \in [q]^n$:*

$$\big|\hat{\mathcal{U}}_i^t(\mathbf{x}) - \bar{u}_i(\mathbf{x})\big| \;\leq\; \hat{b}_i^t(\mathbf{x})$$

where $\hat{\mathcal{U}}_i^t(\mathbf{x})$ is the empirical utility estimate and $\hat{b}_i^t(\mathbf{x})$ is the exploration bonus defined in Eq (4).

Proof. Without loss of generality, assume utilities lie in $[0, 1]$ (the general case follows by scaling with $u_{\max}$). Fix any agent $i \in N$, any coalition label $\ell \in [q]$ (equivalently, any $\mathbf{x}$ with $x_i = \ell$), and any round $t \leq T$. Under bandit feedback, the estimator $\hat{u}_i^t(\mathbf{x})$ is the empirical mean utility, with mean $\bar{u}_i(\mathbf{x})$.

By Hoeffding's inequality (Hoeffding 1963), for any $\hat{b}_i^t(\mathbf{x}) > 0$,

$$\Pr\Big(\big|\hat{\mathcal{U}}_i^t(\mathbf{x}) - \bar{u}_i(\mathbf{x})\big| > \hat{b}_i^t(\mathbf{x})\Big) \leq 2\exp(-2f_{i,\ell}^t(\mathbf{x})(\hat{b}_i^t(\mathbf{x}))^2)$$

Substituting the exploration bonus from Eq (4), we obtain

$$\begin{aligned} &\Pr\Big(\big|\hat{\mathcal{U}}_i^t(\mathbf{x}) - \bar{u}_i(\mathbf{x})\big| > \hat{b}_i^t(\mathbf{x})\Big) \\ &\quad\leq 2\exp\Big(-2n^3 \log\Big(\tfrac{4(n^2+1)T}{\delta}\Big)\Big) \qquad (7) \\ &\quad= 2\left(\tfrac{4(n^2+1)T}{\delta}\right)^{-2n^3}. \end{aligned}$$

There are at most $n \cdot q \cdot T \leq n^2 T$ distinct triples (i, ℓ, t), and hence at most that many corresponding profiles $\mathbf{x}$. Applying the union bound to $\mathcal{O}(1)$, There are at most $nqT \leq n^2 T$ triples (i, ℓ, t). Applying a union bound,

$$\begin{aligned} &\Pr\Big(\exists\, i, \ell, t : \big|\hat{u}_{i,\ell}^{\,t} - \bar{u}_{i,\ell}^{\,t}\big| > \hat{b}_{i,\ell}^{\,t}\Big) \\ &\qquad\leq 2n^2 T \left(\frac{4(n^2+1)T}{\delta}\right)^{-n^3/2} \qquad (8) \end{aligned}$$

where the last inequality uses the fact that the exponent $2n^3 \geq 2$ and $\frac{4(n^2+1)T}{\delta} > 1$.

Thus, with probability at least $1 - \delta$, the inequality

$$\big|\hat{\mathcal{U}}_i^t(\mathbf{x}) - \bar{u}_i(\mathbf{x})\big| \leq \hat{b}_i^t(\mathbf{x})$$

holds simultaneously for all agents i, all rounds t, and all $\mathbf{x} \in [q]^n$. □

Theorem 2 establishes the Nash regret bounds of the UCB−NE algorithm under bandit feedback:

Theorem 2. *Fix a repeated hedonic skill game, UCB−NE algorithm with exploration bonus as in Eq (4). For any $\delta \in (0, 1]$, UCB−NE algorithm with $\varepsilon = \frac{1}{T}$, obtains the following Nash regret bound with probability at least $1 - \delta$ under bandit feedback*

$$\mathcal{R}^T \leq \mathcal{O}(\sqrt{n^3 T \log(n^2 T/\delta)}(\sqrt{n^3} + 1)$$

Proof. Fix a repeated hedonic skill game and run the UCB−NE algorithm with exploration bonus as in Eq (4). Let the time horizon be T, and define the deviation gain in round t as

$$\Delta_t \;=\; \max_{i \in N} \Big(U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) - U_i(\sigma^t) \Big)$$

so that the (cumulative) Nash regret satisfies

$$\mathcal{R}^T \;=\; \sum_{t=1}^{T} \Delta_t$$

By Lemma 2, with probability at least $1 - \delta$ the following event ε holds simultaneously for all agents $i \in N$, all rounds $t \leq T$, and all strategy profiles $\mathbf{x}) \in [q]^n$:

$$\big|\hat{\mathcal{U}}_i^t(\mathbf{x}) - \bar{u}_i(\mathbf{x})\big| \;\leq\; \hat{b}_i^t(\mathbf{x})$$

where $\bar{u}_i(\mathbf{x})$ is the true expected utility and $\hat{b}_i^t(\mathbf{x})$ is the exploration bonus in Eq (4). Under ε, the UCB value $\hat{\mathcal{U}}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$ is a valid upper-confidence bound on $\bar{u}_i(\mathbf{x})$. For $\varepsilon = 1/T$, partition the rounds into

$$\mathcal{T}_{\text{small}} = \{t : \Delta_t \leq \varepsilon\}, \qquad \mathcal{T}_{\text{large}} = \{t : \Delta_t > \varepsilon\}$$

The cumulative regret from small-deviation rounds is bounded by

$$\sum_{t \in \mathcal{T}_{\text{small}}} \Delta_t \;\leq\; |\mathcal{T}_{\text{small}}| \cdot \varepsilon \;\leq\; T \cdot \frac{1}{T} \;=\; 1$$

Thus small rounds contribute only an $\mathcal{O}(1)$ additive term.

Fix any $t \in \mathcal{T}_{\text{large}}$. By definition of Δ_t, there exists an agent i satisfying

$$U_i(\sigma_i^{\star,t}, \sigma_{-i}^t) \;-\; U_i(\sigma^t) \;>\; \varepsilon$$

Under event ε, every coalition $x_i^t \in [q]$ for agent i satisfies

$$\bar{u}_i(\mathbf{x}) \;\leq\; \hat{\mathcal{U}}_i^t(\mathbf{x}) + \hat{b}_i^t(\mathbf{x})$$

Because UCB−NE selects a label whose UCB is maximal, a gap of more than ε between the true best-response payoff and the chosen payoff implies that the chosen action must have had its UCB inflated sufficiently by the exploration

bonus. Thus, for every $t \in \mathcal{T}_{\text{large}}$, there exists an agent i such that

$$\hat{b}_i^t(\mathbf{x}) \;\geq\; \frac{\varepsilon}{2} \tag{9}$$

Recall the bonus form

$$\hat{b}_i^t(\mathbf{x}) = \sqrt{\frac{n^3 \log\big(4(n^2+1)T/\delta\big)}{f_{i,\ell}^T(\mathbf{x}) \vee 1}}$$

From Eq (9), the condition $\hat{b}_i^t(\mathbf{x}) \geq \varepsilon/2$ implies

$$f_{i,\ell}^T(\mathbf{x}) \;\leq\; \frac{4n^3 \log(4(n^2+1)T/\delta)}{\varepsilon^2}$$

Thus each agent–label pair (i, ℓ) can trigger a large-bonus event at most

$$\mathcal{O}\left(\frac{n^3 \log(n^2 T/\delta)}{\varepsilon^2}\right)$$

times. Since there are at most $n \cdot q \leq n^2$ such pairs,

$$|\mathcal{T}_{\text{large}}| \;\leq\; n^2 \cdot \mathcal{O}\left(\frac{n^3 \log(n^2 T/\delta)}{\varepsilon^2}\right) = \mathcal{O}\left(\frac{n^5 \log(n^2 T/\delta)}{\varepsilon^2}\right)$$

The above counting argument is a worst-case bound. To obtain the desired $\sqrt{T}$ dependence, we follow the standard UCB *self-bounding* method: write the regret as

$$\mathcal{R}^T = \sum_{i \in N} \sum_{\ell \in [q]} \Delta_{i,\ell}\, f_{i,\ell}^T(\mathbf{x})$$

where $\Delta_{i,\ell}$ is the (unknown) improvement gap for agent i when switching to label ℓ. Under ε, whenever an agent chooses label ℓ at a round with nonzero gap, we must have $\Delta_{i,\ell} \leq 2\hat{b}_i^t(\mathbf{x})$. Summing over time and applying Cauchy–Schwarz,

$$\sum_{t=1}^{T} \Delta_t \;\leq\; \mathcal{O}\Big(\sqrt{T \log(n^2 T/\delta)}\,\Big(\sqrt{n^3}+1\Big)\Big)$$

where the factor $\sqrt{n^3}$ arises from the scale of the bonus term $\sqrt{n^3 \log(\cdot)/M}$, and the additional $+1$ term accounts for small-gap rounds.

Combining the $\mathcal{O}(1)$ contribution of small-deviation rounds with the refined bound above yields, with probability at least $1-\delta$,

$$\mathcal{R}^T \;\leq\; \mathcal{O}\Big(\sqrt{n^3 T \log(n^2 T/\delta)}\,\Big(\sqrt{n^3}+1\Big)\Big)$$

which establishes the theorem. □

Experimental Results

Dataset: The problem instances used in our experiments are part of the multi-robot task assignment (MRTA)-Benchmark dataset (Bichler, Gimenez, and Alonso-Mora 2025), which includes 250K problems generated in different scenarios with various agents, skills, and tasks.

Baselines and Configurations: To demonstrate how our developed online learning can improve multi-agent planning and coalition formation for maximizing utility of agents (i.e., social welfare) and minimizing Nash regret, we compared UCB−NE with epsilon greedy and Thompson Sampling (Russo and Van Roy 2013) for coalition formation with 500 iterations. For the UCB−NE method, we adopt confidence parameter $\alpha = 0.1$, and reward noise standard deviation $\sigma = 0.1$. The epsilon greedy baseline is configured with an initial exploration rate of $\epsilon = 0.4$, a decay factor of 0.9, and a minimum value of 0.01.

Environment: All experiments were implemented in Python and executed on a system equipped with an Intel Core Ultra 7 processor (3.9GHz, 20 cores), 64GB of RAM, and an NVIDIA GeForce RTX 5070 GPU with 12GB GDDR7 memory.

Sensitivity Analysis: Fig. 1 depicts how varying δ-Confidence Parameter affects the performance of our developed algorithm. This experiment was conducted using five agents, three skills, and 10 tasks. This directly controls the exploration-exploitation tradeoff in Eq. (4). The left-side plot indicates that very small δ values (e.g., 0.01) result in high regret due to overly conservative behavior that limits exploration of better coalition structures. As δ increases, regret decreases sharply and reaches its minimum at $\delta = 0.1$, indicating a more effective balance between exploration and stability. However, larger δ values increase regret as agents pursue uncertain payoffs more aggressively. Thus, a moderate δ offers the lowest regret and the closest progression toward Nash-stable coalition formation. The right plot illustrates the effect of δ on the collective utility. Social welfare generally increases with higher δ values and reaches its maximum at $\delta = 0.8$. Very small values (e.g., around 0.1) hinder convergence due to frequent coalition changes, while moderately larger δ encourages more flexible re-evaluation of partnerships, enabling agents to discover higher-value, complementary coalitions and achieve substantially improved system-wide welfare.

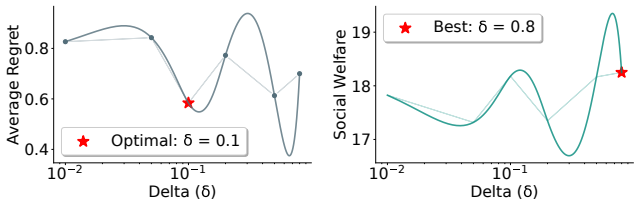

Figure 1: Sensitivity of the UCB−NE algorithm to the exploration parameter δ in terms of regret and social welfare.

Fig. 2 evaluates the scalability of RHSG with UCB−NE, we scale the number of agents $n \in \{5, 10, 15, 20, 50\}$ (left plot), the skill number $k \in [1, 5]$ (middle plot), and the task number $m \in [2, 10]$ (right plot) and compare the Nash regret for different scales of the problems. In agent scalability (#skills=3 and #tasks=10), the plot indicates that the coordination becomes progressively challenging with larger populations. The standard deviation also grows with agent count, reflecting higher variability in system performance. Notably, the final average regret consistently remains below the mean, demonstrating that the algorithm converges to-

ward improved coalition allocations over time. These results highlight that the approach is effective for small to moderate agent populations, while larger systems may require enhanced coordination mechanisms to maintain low regret. For task scalability (#agents=5 and #skills=3), the results indicate that although regret rises, it remains relatively low overall, suggesting multi-agent planning using RHSG still maintains reasonable performance but does not scale optimally as task complexity grows. For skill scalability (#agents=5 and #tasks=10), with increasing the number of skills agents gain more capability combinations, resulting in a larger decision space and more potential assignments to evaluate. This increase in complexity is reflected in the regret values, which capture the learning effort required to identify high-quality allocations. Overall, the UCB−NE algorithm continues to converge toward effective task–skill assignments even as the feasibility constraints and coalition structures become more complex.

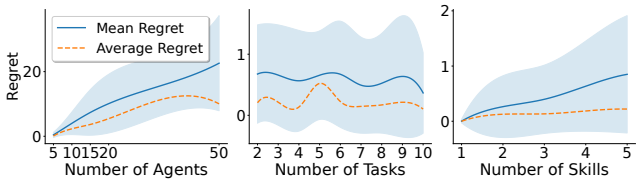

Figure 2: Scalability of the repeated HSG with UCB−NE for different numbers of agents, skills, and tasks.

Comparison Results: Fig. 3 presents a comparative evaluation of UCB-NE, Thompson Sampling, and epsilon greedy on a hedonic skill game instance with 5 agents, 3 skills, and 2 tasks over T=500 rounds in terms of average regret and social welfare per round. The left panel reports the average Nash regret per round. UCB−NE achieves the lowest and smoothest regret trajectory, converging steadily toward the Nash equilibrium baseline, owing to its deterministic confidence-bound-driven exploration that avoids disruptive strategy oscillations. Thompson sampling shows persistently high regret due to independent posterior sampling across agents, where uncoordinated exploration by a single agent degrades the joint outcome. epsilon greedy exhibits high variance because random exploration perturbs the coalition structure for all agents simultaneously.

The social welfare comparison (the right plot) demonstrates an improvement in the coalition formation with the most utilities. Each agent in the UCB−NE algorithm refines its decisions based on historical payoffs and getting familiar with other agents' skills. Although improvements in the epsilon greedy were more significant for the early iterations, the use of online learning algorithm resulted in a huge increase in social welfare. This is because UCB−NE explicitly balances exploration and exploitation when evaluating coalition options. Rather than repeatedly selecting actions based solely on immediate or early payoffs, as in epsilon greedy, UCB−NE incorporates statistical confidence bounds that guide agents to explore promising coalitions that may initially appear suboptimal. Over repeated interactions, each agent accumulates payoff experience while also learning the complementary skills and task contributions of others, allowing coalitions to converge toward more efficient, utility-maximizing structures.

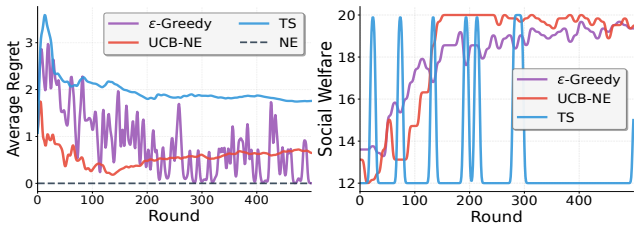

Figure 3: Regret and social welfare for multi-agent planning over the round.

Conclusion and Future Work

This paper studied decentralized coalition formation in *repeated* Hedonic Skill Games (HSGs), where agents repeatedly form coalitions under incomplete information and receive only bandit feedback on realized utilities. This setting captures realistic multi-agent planning scenarios in which agents lack prior knowledge of task requirements, other agents' skills, or counterfactual outcomes. We proposed UCB−NE, a fully decentralized online learning algorithm based on optimism under uncertainty. We showed that UCB−NE achieves sublinear Nash regret and converges with high probability to ε-approximate Nash equilibria, even in general HSGs where pure Nash equilibria may not exist. Importantly, these guarantees are obtained without centralized coordination or inter-agent communication. Empirical evaluations demonstrated that UCB−NE consistently converges to stable coalition structures with higher social welfare than baseline methods, while scaling to problem sizes where equilibrium computation is infeasible. Together, these results establish repeated HSGs as a viable framework for learning-based coalition formation under minimal informational assumptions.

Several directions remain for future investigation. First, our analysis assumes bandit feedback, where agents observe only their realized utilities. Extending the framework to richer feedback models, such as semi-bandit or structured feedback, may yield improved learning efficiency and faster convergence. Second, the current model assumes stationary task values and skill distributions. Studying repeated HSGs in non-stationary environments, where tasks or skills evolve over time, would enable dynamic regret analysis and broaden applicability to adaptive planning domains. Another important direction is to relax the no-communication assumption. Allowing limited or costly communication, such as signaling skills or coalition satisfaction, raises fundamental questions about the trade-off between information exchange and equilibrium convergence. Additionally, the current model assumes no externalities between coalitions. Incorporating inter-coalition externalities arising from shared resources or task interference would connect repeated HSGs to congestion and resource allocation games.

References

Aziz, H.; and Brandl, F. 2012. Existence of stability in hedonic coalition formation games. *arXiv preprint arXiv:1201.4754*.

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.; and Savani, R. 2016. Hedonic games. *Artificial Intelligence*, 235: 27–51.

Bachrach, Y.; Markakis, E.; and Zuckerman, M. 2013. Computing core stable outcomes in skill games. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 26–32.

Bachrach, Y.; and Rosenschein, J. S. 2008. Coalitional skill games. In *7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, 1023–1030.

Bichler, J.; Gimenez, A. M.; and Alonso-Mora, J. 2025. SADCHER: Scheduling using Attention-based Dynamic Coalitions of Heterogeneous Robots in Real-Time. *arXiv preprint arXiv:2510.14851*.

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.

Bogomolnaia, A.; and Jackson, M. O. 2002. Stability in hedonic coalition formation games. *Games and Economic Behavior*, 38(2): 201–230.

Brandt, F.; Bullinger, M.; and Wilczynski, A. 2021. Reaching individually stable coalition structures in hedonic games. In *AAAI Conference on Artificial Intelligence*, volume 35, 5211–5218.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Cohen, G.; and Agmon, N. 2024. Online learning of stable partitions in hedonic games. In *33th International Joint Conference on Artificial Intelligence (IJCAI)*, 1234–1240.

Cohen, S.; and Agmon, N. 2025. Online Learning of Coalition Structures by Selfish Agents. In *AAAI Conference on Artificial Intelligence*, volume 39, 13709–13717.

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.; Goldberg, P. W.; and Papadimitriou, C. H. 2009. The complexity of computing a Nash equilibrium. In *41st annual ACM symposium on Theory of computing*, 89–97.

Ding, Y.; et al. 2022. Independent learning with bandit feedback in multi-agent settings. *Artificial Intelligence*, 301: 103568.

Drèze, J. H.; and Greenberg, J. 1980. Hedonic Coalitions: Optimality and Stability. *Econometrica*, 48(4): 987–1003.

Gourvès, L.; and Monaco, G. 2025. Existence, Computation and Efficiency of Nash Stable Outcomes in Hedonic Skill Games. *Journal of Artificial Intelligence Research*, 82: 1711–1742.

Gourvès, L.; and Monaco, L. 2024. Nash Stability in Hedonic Skill Games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. In *Journal of the American statistical association*, volume 58, 13–30.

Jones, M.; Nguyen, H.; and Nguyen, T. 2023. An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In *AAAI Conference on Artificial Intelligence*, volume 37, 8159–8167.

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

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 *International Conference on Machine Learning*, 7001–7010. PMLR.

Nash Jr, J. F. 1950. Equilibrium points in n-person games. *National academy of sciences*, 36(1): 48–49.

Russo, D.; and Van Roy, B. 2013. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26.

Sliwinski, J.; and Zick, Y. 2017. Learning hedonic games from samples. In *26th International Joint Conference on Artificial Intelligence (IJCAI)*, 391–397.

Valizadeh, J.; Zhang, D.; and Mubin, O. 2024. Enhancing the Efficiency of Systems with Overlapping Coalition Formation. In *Pacific Rim International Conference on Artificial Intelligence*, 284–290. Springer.

Valizadeh, J.; Zhang, D.; and Mubin, O. 2025a. Autonomy with Structural Task Allocation Games: From Inefficiency to Optimality. In *International Conference on Principles and Practice of Multi-Agent Systems*, 435–452. Springer.

Valizadeh, J.; Zhang, D.; and Mubin, O. 2025b. Learning Preferences in Additive Separable Hedonic Project Games. In *Australasian Joint Conference on Artificial Intelligence*, 491–505. Springer.

Von Martial, F. 1992. *Coordinating plans of autonomous agents*. Springer.

Zinkevich, M. 2007. Regret Minimization in Games with Incomplete Information. *20th Annual Conference on Neural Information Processing Systems (NIPS)*, 1729–1736.