Online Learning for Project Selection in Hedonic Project Games

Jaber Valizadeh[1] **Dongmo Zhang**[1] **Omar Mubin**[1] **Ray Telikani**[2]

[1]School of Computing, Data and Mathematical Sciences, Western Sydney University, Australia
[2]Data Science Institute, University of Technology Sydney, Australia

Abstract

We study Hedonic Project Games, a model in which agents select projects with divisible rewards while holding subjective preferences over coalition composition. This framework addresses a key gap in existing models: agents simultaneously care about *who* they collaborate with and *what* they work on. We extend this framework to an online learning setting in which rewards and preferences are initially unknown and learned through repeated interactions. We develop a decentralized variance-reduced stochastic policy-gradient method that provably converges to approximate Nash equilibria using single-sample gradient estimates, and an unbiased momentum variant that achieves faster convergence at the cost of additional curvature computation. We provide sample-complexity guarantees for both methods and show that exploiting the game's additive pairwise structure reduces the interactions required to reach an approximate equilibrium by a polynomial factor in the number of agents. Experiments on synthetic instances and a real-world crowdsourcing dataset validate the theoretical convergence rates and demonstrate the advantages of momentum-based variance reduction over momentum-free baselines.

1 INTRODUCTION

It is the beginning of a new semester in a university research department. The head of the department introduces a set of m new research projects, each associated with its own funding and rewards, and invites n researchers, each of whom can join exactly one project. As you enter the meeting room, you see that several researchers have already chosen some projects. You would like to choose a project that maximizes your own share of the reward while also considering the colleagues involved: you may prefer to collaborate with close friends or perhaps reconnect with colleagues you have not worked with recently. This scenario captures a fundamental tension absent from most existing models: agents care simultaneously about *what* they work on and *who* they work with. Your decision depends not just on the project's reward, but on how that reward is shared and whether your teammates enhance or diminish your experience. How should you choose? And when everyone reasons similarly, what outcomes emerge, and are they efficient?

Existing models typically capture only one side of this tension. The *congestion games* Rosenthal [1973], *project games* Bilò et al. [2023], market sharing games Goemans et al. [2004], *structural task allocation games* Valizadeh et al. [2025a] model competition over rewards, abstracting away from who one collaborates with and implicitly treating teammates as interchangeable. *Hedonic games* Dreze and Greenberg [1980], Bogomolnaia and Jackson [2002], in contrast, model rich preferences over coalition composition but ignore the presence of external, rivalrous rewards tied to specific tasks or projects. *Group activity selection problems* Darmann et al. [2012] enable agents to choose both groups and activities; however, activities are typically labeled without tangible payoffs. Consequently, neither framework captures settings where agents must simultaneously choose what to work on and with whom, balancing collaboration benefits against competition for limited resources.

Although classical hedonic games typically assume static preferences and full information Aziz and et al. [2019], recent research has begun to explore dynamic and more realistic environments Boehmer et al. [2023], Bullinger and Romen [2024]. For example, learning in coalition formation has been studied under the assumption that agents know their preferences in advance Chalkiadakis et al. [2008], Balcan et al. [2015] or within restricted action spaces Chalkiadakis et al. [2011]. However, much of this prior work either presumes that agents are fully aware of their preferences or focuses exclusively on cooperative settings without considering strategic deviations.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6851–6861.

Table 1: Bias–variance–equilibrium in online ASHPG.

Method	Bias	Variance Rate	Nash Gap
Naive-SGD†	Unbiased	$\mathcal{O}(1)$	$\mathcal{O}(T^{-1/2})$
SCG-M (10)	Biased	$\mathcal{O}(t^{-2/3})$	$\mathcal{O}(T^{-1/3})$
1-SFW-M (12)	Unbiased	$\mathcal{O}(t^{-1})$	$\mathcal{O}(T^{-1/2})$

†Naive-SGD is unbiased with non-vanishing variance $\mathcal{O}(1)$, giving an $\mathcal{O}(T^{-1/2})$ Nash gap; the momentum variants trade bias for decaying variance. A self-contained derivation of the Naive-SGD rate is given in the Appendix.

Table 2: Sample complexity for ε-Nash equilibrium

Method	Iterations for ε-NE	Total Regret
SCG Momentum	$\mathcal{O}(n^6/\varepsilon^3)$	$\mathcal{O}(n^3/T^{1/3})$
1-SFW Momentum	$\mathcal{O}(n^4/\varepsilon^2)$	$\mathcal{O}(n^3/\sqrt{T})$

We introduce Hedonic Project Games, unifying competition for divisible project rewards with additive social preferences over coalition structure, and propose the first online learning framework where agents simultaneously learn *what* to work on and *with whom*. The key technical insight is a gradient-alignment property: under symmetric preferences, each agent's utility gradient matches that of a shared potential function, enabling a single variance-reduced stochastic gradient oracle for both regret minimisation and Nash convergence. As summarised in Table 1, the biased SCG-style estimator trades controlled bias for a variance decay of $\mathcal{O}(t^{-2/3})$, yielding an $\mathcal{O}(T^{-1/3})$ Nash gap, while the unbiased 1-SFW-style estimator removes this bias via a stochastic Hessian-vector correction and achieves the faster $\mathcal{O}(T^{-1/2})$ rate at the cost of $\mathcal{O}(nm)$ per-round communication. The resulting sample complexities (Table 2) are $\mathcal{O}(n^6/\varepsilon^3)$ and $\mathcal{O}(n^4/\varepsilon^2)$, derived from the variance bound $\sigma^2 = \mathcal{O}(n^3)$, Lipschitz constant $\mathcal{O}(n)$, and simplex diameter $\mathcal{O}(\sqrt{n})$.

2 RELATED WORKS

Hedonic games have been extensively studied Banerjee et al. [2001], Cechlárová and Romero-Medina [2001], Bogomolnaia and Jackson [2002], with a comprehensive survey provided by Aziz et al. Aziz and Savani [2016]. Several subclasses capture different preference structures, including Additively Separable Hedonic Games (ASHGs) Bogomolnaia and Jackson [2002], Fractional Hedonic Games (FHGs) Aziz et al. [2019], and Hedonic Coalition Nets Elkind and Wooldridge [2009]. Work on effective representations Cechlárová and Romero-Medina [2001], Ballester [2004], Elkind and Wooldridge [2009], Aziz et al. [2019], stable outcome existence Aziz and Brandl [2012], Bilò et al. [2018], Bullinger and Romen [2024], and algorithmic aspects Brandt et al. [2021] is extensive. However, classical hedonic games ignore external factors such as task rewards, and even expressive variants assume agents only care about group composition.

Group Activity Selection Problems (GASPs) Darmann et al. [2012] generalize hedonic games by allowing agents to choose both a group and an activity. For o-GASP with strict preferences, deciding Nash stable assignment existence is NP-complete Darmann [2015]; parameterized complexity is analyzed in Lee and Williams [2017]. However, activities are typically labeled without real payoffs. Project Games Bilò et al. [2023] fall within *monotone valid utility games* Vetta [2002], Augustine et al. [2015], Goemans et al. [2004], Marden and Roughgarden [2010], Marden and Wierman [2013], modeling agents choosing projects with fixed weight contributions. These connect to load balancing Vöcking [2007], congestion games Rosenthal [1973], and hedonic games Dreze and Greenberg [1980], Aziz and Savani [2016]. Project games admit pure Nash equilibria via congestion game reductions, with price of anarchy bounded by 2 Goemans et al. [2004]. However, they assume agents are indifferent to teammates, a limitation in settings where collaboration matters.

However, existing works on online and dynamic hedonic games unrealistically require that the agents' preferences are fully known. Research on PAC learnability in hedonic games attempts to tackle this issue Sliwinski and Zick [2017], Fioravanti et al. [2023]. Unlike our work, they take a *cooperative approach*, aiming to efficiently infer preferences from a limited, fixed number of offline samples. They also assume exact knowledge of preferences before making decisions, limiting practicality, as agents often need time to learn their own preferences from social interactions. In particular, the PAC learning approach is unfit to our *dynamic* setting due to its static nature, requiring a fixed set of preferences that is known in advance.

Bandit algorithms and reinforcement learning techniques have been applied to decentralized matching and coalition formation problems Liu et al. [2021a,b], Valizadeh et al. [2025b], Valizadeh and Telikani [2026], 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, 2025a,b]. 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. We propose a novel framework that addresses these challenges by examining online learning in project selection through the lens of decentralized decision-making by selfish agents. We develop `DOS-PGA`, a decentralized stochastic policy-gradient method in which each agent constructs a local gradient estimator using only bandit observations.

DOS−PGA is tailored to ASHPG because off-the-shelf potential-/congestion-game learners rely on assumptions that fail here. *(i) Non-anonymity:* payoffs depend on co-participant *identities* via $\bar{p}_i(k)$, not aggregate counts, inducing $\Theta(n^2)$ interactions that make count-based estimators [Cui et al., 2022] inefficient. *(ii) Partial observation:* the potential is revealed only through realised co-participation (structured missing data), which our UCB estimators (2)–(6) exploit and generic bandit solvers do not. *(iii) Sample complexity:* treating the m^n joint actions independently costs $\geq \mathcal{O}(n^6/\varepsilon^3)$, whereas exploiting the additive, pairwise structure gives $\mathcal{O}(n^4/\varepsilon^2)$ (Corollary 3), an n^2 gain. We therefore compare against a *pair-feature UCB* baseline, the general bandit potential-game solver of Cohen et al. [2017] with pairwise features (Section 4), which isolates the potential-aligned oracle; PAC-style methods, needing full preference knowledge, cannot be compared fairly online.

3 DEFINITIONS AND NOTATION

We formalise Additive Separable Hedonic Project Games in two stages. Section 3.1 defines the static game, in which agent preferences and project rewards are fully known, and establishes the potential game structure and Nash equilibrium concept. Section 3.2 extends the model to a repeated setting with incomplete information, where preferences are unknown *a priori* and must be learned through co-participation.

3.1 STATIC GAME MODEL

Let $M = \{j_1, j_2, \ldots, j_m\}$ denote a finite set of projects, and let $N = \{1, 2, \ldots, n\}$ be a finite set of agents. Each agent can participate in any project in M, but selects exactly one, and no agent can participate in more than one project simultaneously. Every project $j \in M$ has an associated positive reward, formally defined by the reward function $r : M \to \mathbb{R}_+$, where $r(j)$ denotes the reward of project j. Each agent must select one of the m projects. Let $s_i \in M$ denote the project chosen by agent i as her *strategy*. A *strategy profile* is a vector $\mathbf{s} = (s_1, \ldots, s_n)$, and the set of all strategy profiles is $\mathcal{S} = M^n$. Given a strategy profile $\mathbf{s} \in \mathcal{S}$, the *coalition* assigned to each project $j \in M$ is denoted by $C_j(\mathbf{s}) = \{i \in N \mid s_i = j\}$, and we write $|C_j(\mathbf{s})|$ for its size.

Each agent i has a (cardinal) *preference* over other agents, encoded by a function $p_i : N \to \mathbb{R}$, where $p_i(k)$ represents the value that agent i assigns to being in the same coalition as agent k, such that $p_i(i) = 0$. Positive values represent attraction (*friends*), negative values represent aversion (*enemies*). Preferences are said to be *symmetric* if $p_i(k) = p_k(i)$ for all $i, k \in N$.

Let $\gamma \geq 0$ be a universal normalizing coefficient to reconcile the units of the reward (e.g. dollars) and the preference score (e.g. utility points). Project rewards are assumed to be divisible and are shared equally among all agents who choose the same project. We assume γ is fixed and common intensity parameter. The *utility* of agent i under strategy profile $\mathbf{s}$ is given by:

$$u_i(\mathbf{s}) = \frac{r(s_i)}{|C_{s_i}(\mathbf{s})|} + \gamma \sum_{\substack{k \in C_{s_i}(\mathbf{s}) \\ k \neq i}} p_i(k) \tag{1}$$

where the first term ensures equal sharing of project rewards[1], and the second term captures additive social preferences, scaled by γ.

An additive separable hedonic project game is formally defined as follows:

Definition 1. *An additive separable hedonic project game (ASHPG) is specified by* $\mathcal{G} = \langle N, M, r, \mathcal{S}, \gamma, (p_i)_{i \in N}, (u_i)_{i \in N} \rangle$*, where the utility of each agent* $i \in N$ *is defined as (1).*

Connections to Existing Models. Our framework generalizes two widely studied models:

- If $p_i(k) = 0$ for all $i, k \in N$, the model reduces to a *symmetric project game* with identical weights Bilò et al. [2023], where agent utilities depend only on the project reward and coalition sizes.
- If $r(j) = 0$ for all $j \in M$, the model reduces to an *additively separable hedonic game* Bogomolnaia and Jackson [2002], where agents care only about which other agents are in their coalition.

Eq. (1) models fixed-budget crowdsourcing platforms such as GMISSION Chen et al. [2014]: each task j carries a *fixed* payment $r(j)$ split among its workers, so a worker's share is the divisible-reward term $r(s_i)/|C_{s_i}(\mathbf{s})|$ (adding a worker dilutes everyone's pay), while spatial-affinity preferences $\bar{p}_i(k)$ (e.g. co-location similarity) form the additive social term. The equal-split rule for a fixed pool is standard in congestion-style games Bilò et al. [2023], Vöcking [2007], Goemans et al. [2004] and is what makes the reward term a harmonic potential (Lemma 1). Size-*dependent* rewards would violate Lemma 1 and define a different model class (future work).

A central question in this setting is whether the project selection process converges to a stable outcome, that is, a *Nash equilibrium*. In our setting, a *deviation* occurs when a player i changes her strategy from the current project s_i under the strategy profile $\mathbf{s}$ to a distinct project $j \neq s_i$. Given a pure-strategy profile $\mathbf{s}$, agent i is said to have a *profitable deviation* if switching from her current project s_i

[1] Since each agent selects a project, for any agent i, $|C_{s_i}(\mathbf{s})| \geq 1$, as $i \in C_{s_i}(\mathbf{s})$.

to an alternative project $j \neq s_i$, while keeping the others' strategies fixed, improves her utility.

A strategy profile $\mathbf{s}$ is *pure* Nash equilibrium if, for all players $i \in N$ and all $j \in M$, it holds

$$u_i(\mathbf{s}) \geq u_i(j, \mathbf{s}_{-i})$$

where $(j, \mathbf{s}_{-i})$ denotes the profile obtained from $\mathbf{s}$, after player i deviates from s_i to j.

For each player i, the strategy space is M. Let $\Delta_i = \Delta(M) = \left\{\pi_i \in [0,1]^m : \sum_{j \in M} \pi_i(j) = 1\right\}$ be the policy space is the probability simplex over projects, for each player $i \in N$. The joint policy space is $\Delta = \prod_{i \in N} \Delta_i$. A joint policy $\pi = (\pi_1, \ldots, \pi_n)$ is defined as an element of Δ, where $\mathbf{s} = (s_1, \ldots, s_n) \sim \pi$ denotes that $s_i \sim \pi_i$ independently across all $i \in N$. Given a joint policy $\pi \in \Delta$, the expected utility of each agent i is denoted by

$$U_i^{\pi} = \mathbb{E}_{\mathbf{s} \sim \pi}[u_i(\mathbf{s})] = \sum_{\mathbf{s} \in \mathcal{S}} \Big(\prod_{\ell \in N} \pi_\ell(s_\ell) \Big) u_i(\mathbf{s})$$

Given a general policy π, let π_{-i} denote the marginal joint policy of players $1, \ldots, i-1, i+1, \ldots, n$. The best response to player i under policy π is defined as $\pi_i^{\dagger} = \arg\max_{\mu \in \Delta(M)} U_i^{\mu, \pi_{-i}}$, and the corresponding expected utility is $U_i^{\dagger, \pi_{-i}} = U_i^{\pi_i^{\dagger}, \pi_{-i}}$.

Our goal is to compute an approximate Nash equilibrium of the game, defined as follows. A joint policy π is an ε-approximate (mixed) Nash equilibrium if

$$\max_{i \in N} \left(U_i^{\dagger, \pi_{-i}} - U_i^{\pi} \right) \leq \varepsilon$$

Remark 1. *As an exact potential game under symmetric preferences (Lemma 1), ASHPG admits* at least one *pure Nash equilibrium (any pure maximiser of $\tilde{\Phi}^t$). Computing one exactly is intractable and needs full preference knowledge, so our online algorithm instead targets an ε-approximate* mixed *equilibrium via the time-averaged iterate $\bar{\pi}^T = \frac{1}{T} \sum_t \pi^t$ (Theorem 2).*

In many practical scenarios, such as crowdsourcing platforms or distributed multi-agent systems, agents may lack prior knowledge about others' identities or preferences. These preferences are gradually learned through repeated interactions.

3.2 ONLINE LEARNING MODEL

We now extend the additive separable hedonic project game to a repeated setting with incomplete information, in which agents initially lack knowledge of other agents' preferences, and learn them through interaction.

Time proceeds in discrete rounds $t = 1, 2, \ldots, T$. In each round, every agent selects exactly one project and a coalition structure is induced. Recall that preferences are unknown, and even the agents themselves may not be aware of them. Thus, for any pair of distinct agents $i, k \in N$, agent i has an unknown and fixed preference distribution $\mathcal{D}_{i,k}$ supported on $[p_{\min}, p_{\max}] \subset \mathbb{R}$, with mean $\bar{p}_i(k)$ for each agent i who aims to learn, where $\bar{p}_i(i) = 0$. These distributions are assumed to be stationary and independent across time. We assume symmetry of true mean preferences, i.e., $\bar{p}_i(k) = \bar{p}_k(i)$ for all $i, k \in N$, while allowing individual realizations to differ.

At round t, if agents i and k select the same project, agent i observes a realization $p_i^t(k)$ which is then independently drawn from $\mathcal{D}_{i,k}$. We use the convention that $p_i^t(i) = \bar{p}_i(i) = 0$ for any agent i. If they do not co-participate, no observation is obtained. Thus, preference information is revealed only through joint participation. Given a pure strategy profile $\mathbf{s} \in M^n$, the *true utility* of agent i is $\bar{u}_i(\mathbf{s}) = \frac{r(s_i)}{|C_{s_i}(\mathbf{s})|} + \gamma \sum_{\substack{k \in C_{s_i}(\mathbf{s}) \\ k \neq i}} \bar{p}_i(k)$. At time step t, agent i's expected utility under a mixed strategy profile π^t is defined as $U_i^{\pi^t} = \mathbb{E}_{\mathbf{s} \sim \pi^t}[\bar{u}_i(\mathbf{s})]$.

Regret Minimization. To evaluate learning performance, we adopt the standard notion of *regret* Lattimore and Szepesvári [2020] with respect to estimated utilities. Regret quantifies the cumulative difference between the expected utility an agent could have achieved by playing her best response at each round and the expected utility she actually obtained. An ε-approximate Nash equilibrium can be obtained by achieving sublinear Nash regret, defined below.

Definition 2. *Let π^t be the policy at the t-th round. The Nash regret after T round are defined as*

$$\mathcal{R}^T = \sum_{t=1}^{T} \max_{i \in N} \left(U_i^{\dagger, \pi_{-i}^t} - U_i^{\pi^t} \right)$$

A learning algorithm achieves *sublinear regret* if $\mathcal{R}^T / T \to 0$ as $T \to \infty$. We refine the observation model by distinguishing two types of feedback, mirroring the semi-bandit and bandit settings of congestion games Cui et al. [2022].

Definition 3. *At each round t, given the realised strategy profile $\mathbf{s} \sim \pi^t$:*

- ***Semi-bandit feedback:** Agent i observes the individual preference realisations $p_i^t(k)$ for every co-participant $k \in C_{s_i}^t(\mathbf{s}) \setminus \{i\}$, as well as her equal shared project reward $\frac{r(s_i)}{|C_{s_i}^t(\mathbf{s})|}$.*
- ***Bandit feedback:** Agent i observes only her total realised utility $u_i^t(\mathbf{s})$, with no decomposition into reward and preference components.*

Under either type, co-participation is necessary for preference information to be revealed. If agents i and k are never co-located at any project up to round t, the pair preference $\bar{p}_i(k)$ remains entirely unobserved.

3.2.1 Optimism Principle and Feedback Settings

In our setting, agents select projects in an *uncertain* environment where preferences over coalition members are initially *unknown*. To navigate this uncertainty, we introduce a UCB-based algorithm grounded in the principle of *optimism in the face of uncertainty*, which prescribes acting as if the environment is as favorable as plausibly possible Lattimore and Szepesvári [2020].

Semi-Bandit Feedback. Under semi-bandit feedback, agent i maintains an empirical estimate of her preferences based on observed feedback based on co-participation history. Let $N_{i,k}^t(\mathbf{s}) = \sum_{\tau=1}^t \mathbf{1}\{s_i^\tau = s_k^\tau\}$, denote the number of rounds up to time t in which agents i and k co-participated. The empirical estimate of agent i's preference for agent k at time t is defined as

$$\hat{p}_i^t(k) = \begin{cases} \dfrac{\sum_{\tau=1}^t p_i^\tau(k)\,\mathbf{1}\{s_i^\tau = s_k^\tau\}}{N_{i,k}^t(\mathbf{s})} & \text{if } N_{i,k}^t(\mathbf{s}) > 0, \\ p_{\max} & \text{if } N_{i,k}^t(\mathbf{s}) = 0. \end{cases} \tag{2}$$

The exploration bonus is

$$\hat{\beta}_{i,k}^t(\mathbf{s}) = \sqrt{\frac{2\log(4n^2T/\delta)}{N_{i,k}^t(\mathbf{s}) \vee 1}}$$

where, $\delta \in (0,1]$ is a confidence parameter controlling the exploration-exploitation trade-off, and $a \vee b = \max\{a,b\}$. At round t, for each agent i, and $k \in C_{s_i}^t(\mathbf{s})$, the optimistic preference estimate is $\tilde{p}_i^t(k) = \hat{p}_i^t(k) + \hat{\beta}_{i,k}^t(\mathbf{s})$.

Based on optimistic estimates, agent i evaluates outcomes using the *optimistic utility*:

$$\tilde{u}_i^t(\mathbf{s}) = \frac{r(s_i)}{|C_{s_i}^t(\mathbf{s})|} + \gamma \sum_{\substack{k \in C_{s_i}^t(\mathbf{s}) \\ k \neq i}} \tilde{p}_i^t(k) \tag{3}$$

Bandit Feedback. Under bandit feedback, agent i observes only her total realised utility $u_i^t(\mathbf{s})$ at each round, with no decomposition into the reward and preference components. For each project $j \in M$, let $M_{i,j}^t(\mathbf{s}) = \sum_{\tau=1}^t \mathbf{1}\{s_i^\tau = j\}$ denote the number of rounds up to time t in which agent i selected project j under profile $\mathbf{s}$. The empirical estimate of agent i's utility from project j at time t is defined as

$$\check{u}_i^t(j, \mathbf{s}_{-i}) = \begin{cases} \dfrac{\sum_{\tau=1}^t u_i^\tau(\mathbf{s})\,\mathbf{1}\{s_i^\tau = j\}}{M_{i,j}^t(\mathbf{s})} & \text{if } M_{i,j}^t(\mathbf{s}) > 0, \\ u_{\max} & \text{if } M_{i,j}^t(\mathbf{s}) = 0 \end{cases} \tag{4}$$

where $u_{\max} = r_{\max} + \gamma(n-1)p_{\max}$ is an upper bound on agent utility. The exploration bonus is

$$\check{\beta}_{i,j}^t(\mathbf{s}) = \sqrt{\frac{2\log(2mT/\delta)}{M_{i,j}^t(\mathbf{s}) \vee 1}} \tag{5}$$

and the optimistic project utility estimate for agent i is

$$\tilde{\mathcal{U}}_i(j, \mathbf{s}_{-i}) = \check{u}_i^t(j, \mathbf{s}_{-i}) + \check{\beta}_{i,j}^t(\mathbf{s}) \tag{6}$$

3.3 EXISTENCE OF A NASH EQUILIBRIUM

A key property enabling our approach is that ASHPG admits a potential game structure under symmetric preferences.

Lemma 1. *For each round t, under symmetric preferences ($\bar{p}_i(k) = \bar{p}_k(i)$ for all $i,k \in N$), the game $\mathcal{G}^t = \langle N, M, r, \mathcal{S}, \gamma, (\tilde{p}_i^t), (\tilde{u}_i^t)\rangle$ is an exact potential game with potential function*

$$\tilde{\Phi}^t(\mathbf{s}) = \sum_{j \in M} \sum_{Z=1}^{|C_j^t(\mathbf{s})|} \frac{r(j)}{Z} + \frac{\gamma}{2} \sum_{j \in M} \sum_{\substack{i,k \in C_j^t(\mathbf{s}) \\ k \neq i}} \tilde{p}_i^t(k)$$

Let $\tilde{\Phi}(\pi) = \mathbb{E}_{\mathbf{s}\sim\pi}[\tilde{\Phi}(\mathbf{s})]$ denote the multilinear extension of the potential function. For each agent i and project j:

$$\frac{\partial \tilde{\Phi}(\pi)}{\partial \pi_i(j)} = \mathbb{E}_{\mathbf{s}_{-i}\sim\pi_{-i}}\left[\tilde{\Phi}(j, \mathbf{s}_{-i}) - \tilde{\Phi}(\mathbf{s}_{-i})\right]$$

Using the exact potential property (Lemma 1),

$$\tilde{\Phi}(j, \mathbf{s}_{-i}) - \tilde{\Phi}(\mathbf{s}_{-i}) = u_i(j, \mathbf{s}_{-i}) - u_i(\mathbf{s}_{-i})$$

which implies

$$\nabla_{\pi_i} \tilde{\Phi}(\pi) = \nabla_{\pi_i} U_i^\pi, \qquad \forall\, i \in N \tag{7}$$

Lemma 2. *The expected potential $\Phi : \Delta \to \mathbb{R}$ has L-Lipschitz continuous gradients with*

$$L = 2r_{\max} + 2\gamma(n-1)p_{\max} \tag{8}$$

where $r_{\max} = \max_j r(j)$ and $p_{\max} = \max_{i \neq k} |\bar{p}_i(k)|$.

Given a joint policy π, define the Nash gap by $\mathsf{NG}(\pi) := \max_{i \in N}\left(U_i^{\dagger,\pi_{-i}} - U_i^\pi\right)$. By the gradient alignment (7), the per-agent best-response gain equals the Frank–Wolfe gap of Φ with respect to agent i's simplex direction.

We now establish the connection between gradient norms and approximate Nash equilibria.

Corollary 1. *Using the gradient alignment* (7)*, we have*

$$\mathsf{NG}(\pi) \leq 2 \max_{i \in N} \|\nabla_{\pi_i} \Phi(\pi)\|_\infty$$

Equality holds when, for some agent $i^\star$, $\pi_{i^\star}$ is concentrated on the project with the minimum *gradient component, so the weighted average of gradients equals the minimum.*

Beyond symmetry. Symmetry ($\bar{p}_i(k) = \bar{p}_k(i)$) is what makes $\tilde{\Phi}^t$ an *exact* potential. With asymmetry level $\delta := \max_{i,k} |\bar{p}_i(k) - \bar{p}_k(i)|$, the potential Φ built from symmetrised preferences is an $\mathcal{O}(\delta)$-approximate potential (gradient mismatch $\leq \gamma(n-1)\delta$), and `DOS-PGA` still converges to a stationary point with $\mathsf{NG}(\pi) \leq 2\max_i \|\nabla_{\pi_i}\Phi(\pi)\|_\infty + \mathcal{O}(\delta)$, recovering Corollary 1 when $\delta = 0$ (Corollary 4, appendix).

3.4 DISTRIBUTED ONE-SAMPLE ALGORITHM

We present our main algorithm, which combines the potential game structure with variance-reduced stochastic gradient methods Mokhtari et al. [2018, 2020], Zhang et al. [2020]. Algorithm 1, that we called DOS−PGA (Distributed One-Sample Projected Gradient Ascent), presents a variance-reduced stochastic projected gradient ascent method for online learning in ASHPGs.

The algorithm offers two variants:

- **Option A (Biased, SCG-style):** Achieves $\mathcal{O}(T^{-1/3})$ Nash gap with minimal per-iteration cost. It is *fully decentralized with zero communication*: each agent updates using only its own bandit observations.
- **Option B (Unbiased, 1-SFW-style):** Incorporates an additional correction term $\tilde{\mathbf{\Delta}}_i^t$ that estimates the gradient drift between consecutive iterates, restoring unbiasedness and improving the convergence rate to $\mathcal{O}(T^{-1/2})$ (which is faster) at the expense of computing a stochastic Hessian-vector product.

Communication cost. Option A requires *no* communication. Option B broadcasts only each agent's *strategy update* $\pi_i^t - \pi_i^{t-1} \in \mathbb{R}^m$ per round (not raw preferences, which stay private), an $\mathcal{O}(nm)$ payload[2] plus an $\mathcal{O}(nm)$ per-iteration compute overhead from the Jacobian-vector product (15). This makes the speed–overhead trade-off explicit (Option A for constrained settings, Option B for the faster $\mathcal{O}(T^{-1/2})$ rate); since B's per-round overhead is constant while its gap decays faster, a crossover round governs the wall-clock-vs-Nash-gap comparison.

3.4.1 Stochastic Gradient Estimation

Computing the exact gradient $\nabla_{\pi_i}\Phi(\pi^t)$ is infeasible for two reasons: (i) the true mean preferences $\bar{p}_i(k)$ are unknown, and (ii) the expectation involves a sum over m^n strategy profiles. Instead, each agent constructs a *one-sample* stochastic gradient estimate using only the information revealed in round t.

In round t, given sampled profile $\mathbf{s}$ and observed preferences $\{p_i^t(k)\}_{k\in C_{s_i}^t(\mathbf{s})}$, agent i constructs $\widehat{\nabla}_{\pi_i}\Phi(\pi^t) \in \mathbb{R}^m$ with components:

$$\left[\widehat{\nabla}_{\pi_i}\Phi(\pi^t)\right]_j = \widehat{\Phi}_i^t(j, \mathbf{s}_{-i}) - \widehat{\Phi}_i^t(\mathbf{s}_{-i}), \qquad \forall\, j \in M \quad (9)$$

where $\widehat{\Phi}_i^t(j, \mathbf{s}_{-i})$ is the estimated potential when agent i plays j against the realised opponents' profile $\mathbf{s}_{-i}$, and $\widehat{\Phi}_i^t(\mathbf{s}_{-i})$ is the estimated *marginal* potential of the opponents' coalition (i.e. the potential evaluated without agent i's contribution). This ensures the estimator is centred correctly around $\nabla_{\pi_i}\Phi(\pi^t)$.

[2] At our largest scale (n=100, m=200, double precision), $\approx 100 \times 200 \times 8\,\mathrm{B} = 160\,\mathrm{KB}$ per round, within standard multi-agent budgets.

Algorithm 1 DOS−PGA for ASHPG

Input: Time horizon T, step size $\eta_t = 1/T$, averaging parameter $\rho_t = 4/(t+8)^{2/3}$

Initialize $\pi_i^1 = \mathbf{1}/m$ uniformly for all $i \in N$; set $\mathbf{d}_i^0 = \mathbf{0}$
for $t = 1, 2, \ldots, T$ **do**
  Each agent i samples project $s_i \sim \pi_i^t$
  Form strategy profile $\mathbf{s} = (s_1, \ldots, s_n)$ and induce coalitions $\{C_j^t(\mathbf{s})\}_{j\in M}$
  Each agent i observes preferences $\{p_i^t(k) : k \in C_{s_i}^t(\mathbf{s}) \setminus \{i\}\}$ and computes realized utility $\tilde{u}_i^t(\mathbf{s})$ via (3)
  Compute stochastic gradient estimate $\widehat{\nabla}_{\pi_i}^t\Phi(\pi)$ via (9)
  Option A (Biased): $\mathbf{d}_i^t = (1-\rho_t)\mathbf{d}_i^{t-1} + \rho_t \widehat{\nabla}_{\pi_i}^t\Phi(\pi)$
  Option B (Unbiased): Sample $\alpha \sim \mathrm{Uniform}[0,1]$, compute $\tilde{\mathbf{\Delta}}_i^t$ via (15)
    $\mathbf{d}_i^t = (1-\rho_t)(\mathbf{d}_i^{t-1} + \tilde{\mathbf{\Delta}}_i^t) + \rho_t \widehat{\nabla}_{\pi_i}^t\Phi(\pi)$
  Update mixed strategy: $\pi_i^{t+1} = \Pi_{\Delta(M)}\left[\pi_i^t + \eta_t \mathbf{d}_i^t\right]$
end for
return $\bar{\pi}^T = \frac{1}{T}\sum_{t=1}^T \pi^t$

Estimator (9) **under limited feedback.** No importance weighting is needed: actions are sampled on-policy and every unknown is replaced by an *optimistic* (UCB) estimate, so optimism plays the role of off-policy correction. The feedback models of Definition 3 differ only in how $\widehat{\Phi}_i^t$ is formed. *Semi-bandit:* the per-pair UCB estimates $\tilde{p}_i^t(k)$ of (2), with fallback $p_{\max}$ for un-observed pairs. *Bandit:* the per-project UCB utilities $\tilde{\mathcal{U}}_i^t(j, \mathbf{s}_{-i})$ of (6), which cannot separate reward-sharing from preferences; this inflates σ^2 by a $\log T$ factor but leaves every rate *exponent* in Theorems 1–3 unchanged, so the $\mathcal{O}(T^{-1/3})$ and $\mathcal{O}(T^{-1/2})$ Nash-gap rates hold under both models.

A naive single-sample estimator used directly for gradient ascent has variance that does not diminish over time, leading to $\mathcal{O}(1/\sqrt{T})$ convergence. To achieve faster convergence we employ the momentum-based variance reduction techniques described below.

3.4.2 Biased Momentum Estimator (SCG-style)

We adopt the momentum-based averaging technique from the Stochastic Continuous Greedy (SCG) framework Mokhtari et al. [2018]. Agent i updates her gradient estimate as:

$$\mathbf{d}_i^t = (1-\rho_t)\,\mathbf{d}_i^{t-1} + \rho_t\,\widehat{\nabla}_{\pi_i}\Phi(\pi^t) \qquad (10)$$

where $\rho_t = \frac{4}{(t+8)^{2/3}}$, initialised with $\mathbf{d}_i^0 = \mathbf{0}$.

This recursion implements an exponential moving average. The specific choice $\rho_t = \mathcal{O}(t^{-2/3})$ decays slowly enough to incorporate new gradient information as π^t evolves, yet fast enough to ensure the estimation variance vanishes over time.

Assumption 1 (Gradient Smoothness)**.** *The expected potential function $\Phi : \Delta \to \mathbb{R}$ has L-Lipschitz continuous gradients, i.e., for all $\pi, \pi' \in \Delta$:*

$$\|\nabla\Phi(\pi) - \nabla\Phi(\pi')\| \le L\|\pi - \pi'\|$$

where $L = 2r_{\max} + 2\gamma(n-1)p_{\max}$ as derived in Lemma 2.

This estimator is *biased*: since $\mathbf{d}_i^{t-1}$ approximates $\nabla_{\pi_i}\Phi(\pi^{t-1})$ rather than $\nabla_{\pi_i}\Phi(\pi^t)$, we have $\mathbb{E}[\mathbf{d}_i^t] \neq \nabla_{\pi_i}\Phi(\pi^t)$ in general. However, because $\eta_t = 1/T$ is small and Φ has L-Lipschitz gradients (Assumption 1), consecutive iterates satisfy $\|\pi^t - \pi^{t-1}\| = \mathcal{O}(1/T)$, which ensures the bias remains controlled.

Lemma 3. *Under Assumption 1, the biased momentum estimator* (10) *satisfies:*

$$\mathbb{E}\big[\|\mathbf{d}_i^t - \nabla_{\pi_i}\Phi(\pi^t)\|^2\big] \le \frac{C_1\sigma^2}{(t+8)^{2/3}} + \frac{C_2 L^2}{T^2} \qquad (11)$$

where $\sigma^2 = \mathbb{E}[\|\widehat{\nabla}_{\pi_i}\Phi(\pi^t) - \nabla_{\pi_i}\Phi(\pi^t)\|^2]$ is the variance of the stochastic gradient, L is the Lipschitz constant from Assumption 1, and $C_1, C_2 > 0$ are universal constants.

3.4.3 Unbiased Momentum Estimator (1-SFW-style)

Although the SCG-style estimator achieves a reduction in variance, its inherent bias limits convergence to $\mathcal{O}(T^{-1/3})$. To obtain the faster rate $\mathcal{O}(T^{-1/2})$ we adapt the unbiased momentum technique from the One-Sample Stochastic Frank-Wolfe (1-SFW) algorithm Zhang et al. [2020].

The key insight is that the bias in (10) arises because $\mathbf{d}_i^{t-1}$ estimates the *previous* gradient $\nabla_{\pi_i}\Phi(\pi^{t-1})$ rather than the *current* one. We construct a correction term using the gradient drift $\boldsymbol{\Delta}_i^t = \nabla_{\pi_i}\Phi(\pi^t) - \nabla_{\pi_i}\Phi(\pi^{t-1})$, yielding:

$$\mathbf{d}_i^t = (1-\rho_t)\big(\mathbf{d}_i^{t-1} + \tilde{\boldsymbol{\Delta}}_i^t\big) + \rho_t\,\widehat{\nabla}_{\pi_i}\Phi(\pi^t) \qquad (12)$$

where $\tilde{\boldsymbol{\Delta}}_i^t$ is an unbiased estimator of $\boldsymbol{\Delta}_i^t$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space. Define the natural filtration $\{\mathcal{F}_t\}_{t\ge 0}$ by

$$\mathcal{F}_t = \sigma\big(\mathbf{s}^1,\, \{p_i^\tau(k)\}_{\tau\le t},\, \{\mathbf{d}_i^\tau\}_{\tau\le t},\, \{\pi^\tau\}_{\tau\le t} :\; i,k \in N\big) \qquad (13)$$

By construction, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$ is an increasing sequence of σ-algebras (a filtration). In particular, π^t and $\mathbf{d}_i^{t-1}$ are both $\mathcal{F}_{t-1}$-measurable, since they are fully determined by observations up to and including round $t-1$.

Proposition 1. *Let $t \ge 2$ and fix agent $i \in N$. Suppose the following three conditions hold:*

1. *$\mathbf{d}_i^{t-1}$ is an unbiased estimator of $\nabla_{\pi_i}\Phi(\pi^{t-1})$, i.e. $\mathbb{E}[\mathbf{d}_i^{t-1} \mid \mathcal{F}_{t-2}] = \nabla_{\pi_i}\Phi(\pi^{t-1})$.*
2. *$\tilde{\boldsymbol{\Delta}}_i^t$ is an unbiased estimator of $\boldsymbol{\Delta}_i^t$, i.e. $\mathbb{E}[\tilde{\boldsymbol{\Delta}}_i^t \mid \mathcal{F}_{t-1}] = \boldsymbol{\Delta}_i^t$.*
3. *$\widehat{\nabla}_{\pi_i}\Phi(\pi^t)$ is an unbiased estimator of $\nabla_{\pi_i}\Phi(\pi^t)$, i.e. $\mathbb{E}[\widehat{\nabla}_{\pi_i}\Phi(\pi^t) \mid \mathcal{F}_{t-1}] = \nabla_{\pi_i}\Phi(\pi^t)$.*

Then $\mathbf{d}_i^t$ defined by (12) *is an unbiased estimator of $\nabla_{\pi_i}\Phi(\pi^t)$.*

By the fundamental theorem of calculus:

$$\boldsymbol{\Delta}_i^t = \int_0^1 \nabla_\pi(\nabla_{\pi_i}\Phi)\big(\pi^{t-1} + \alpha(\pi^t - \pi^{t-1})\big)\cdot(\pi^t - \pi^{t-1})\,d\alpha \qquad (14)$$

where the Jacobian $\nabla_\pi(\nabla_{\pi_i}\Phi) \in \mathbb{R}^{m\times nm}$ is the derivative of the i-th gradient block with respect to the *full* joint strategy vector $\pi \in \mathbb{R}^{nm}$ (capturing the cross-agent terms arising from the multilinear extension).

An unbiased estimator is obtained by randomising the integration point:

1. Sample $\alpha \sim \mathrm{Uniform}[0,1]$.
2. Compute the interpolated strategy $\pi^t(\alpha) = \pi^{t-1} + \alpha(\pi^t - \pi^{t-1})$.
3. Form the stochastic Jacobian-vector product:

$$\tilde{\boldsymbol{\Delta}}_i^t = \widehat{\nabla}_\pi(\nabla_{\pi_i}\Phi)(\pi^t(\alpha))\cdot(\pi^t - \pi^{t-1}) \qquad (15)$$

where $\widehat{\nabla}_\pi(\nabla_{\pi_i}\Phi)$ denotes a stochastic estimate of the full cross-agent Jacobian block at $\pi^t(\alpha)$. Since $\mathbb{E}_\alpha[\nabla_\pi(\nabla_{\pi_i}\Phi)(\pi^t(\alpha))] = \int_0^1 \nabla_\pi(\nabla_{\pi_i}\Phi)(\pi^t(\alpha))\,d\alpha$, this yields $\mathbb{E}[\tilde{\boldsymbol{\Delta}}_i^t] = \boldsymbol{\Delta}_i^t$ as required by assumption (ii) of Proposition 1.

3.4.4 Convergence Analysis

We now establish convergence properties of Algorithm 1. Our analysis proceeds in three stages: (i) bounding the gradient estimation error, (ii) establishing potential function improvement, and (iii) translating potential gaps to Nash gaps via the potential game structure.

Assumption 2 (Bounded Strategy Space)**.** *The mixed strategy space $\Delta = \prod_{i\in N}\Delta(M)$ is compact with diameter D: $\|\pi - \pi'\| \le D$ for all $\pi, \pi' \in \Delta$. For the probability simplex over m projects and n agents, $D = \sqrt{2n}$.*

Assumption 3 (Bounded Gradient Variance)**.** *The stochastic gradient estimator $\widehat{\nabla}_{\pi_i}\Phi(\pi^t)$ has bounded variance: for any $\pi^t \in \Delta$ and $i \in N$,*

$$\mathbb{E}\big[\|\widehat{\nabla}_{\pi_i}\Phi(\pi^t) - \nabla_{\pi_i}\Phi(\pi^t)\|^2 \mid \pi^t\big] \le \sigma^2$$

where the expectation is over the randomness in strategy realisations $\mathbf{s} \sim \pi^t$ *and preference observations* $p_i^t(k) \sim \mathcal{D}_{i,k}$.

Assumption 4 (Bounded Preferences)**.** *The preference distributions have bounded support:* $p_{\min} \le p_i^t(k) \le p_{\max}$ *a.s. for all* $i, k \in N$ *and* $t \ge 1$.

Lemma 4. *Under Assumptions 2 and 4, the gradient variance satisfies:*

$$\sigma^2 \le 4m\big[r_{\max}^2 + \gamma^2(n-1)^2 p_{\max}^2\big] \tag{16}$$

Gradient Estimation Error Analysis. We first establish recursive bounds on the gradient estimation error, which is the key technical ingredient for our convergence analysis.

Potential Function Convergence. We now establish convergence in terms of the expected potential gap.

Assumption 5 (Concavity of Multilinear Extension)**.** *The multilinear extension* $\Phi : \Delta \to \mathbb{R}$ *is concave on* Δ.

Remark 2. *Assumption 5 is stated for reference but* does not hold for ASHPG in general. *By the gradient-alignment identity* (7)*,* $\partial\Phi/\partial\pi_i(j) = \mathbb{E}_{\mathbf{s}_{-i}}[u_i(j, \mathbf{s}_{-i})]$ *is independent of* π_i*, so the diagonal Hessian blocks vanish,* $\partial^2\Phi/\partial\pi_i(j)\,\partial\pi_i(\ell) = 0$ *(see the proof of Lemma 2). A nonzero symmetric matrix with an all-zero diagonal has trace* 0 *and is therefore* indefinite*; hence* Φ *is neither concave nor convex on* Δ *unless* $\nabla^2\Phi \equiv 0$*; a non-negative-preference witness (as in* GMISSION*) is given in Example 1 (appendix).*

Consequently we *do not rely* on Assumption 5: our Nash guarantees rest on the *unconditional* bound of Corollary 1, $\mathsf{NG}(\pi) \le 2\max_i \|\nabla_{\pi_i}\Phi(\pi)\|_\infty$, together with convergence of the per-agent Frank–Wolfe gradient gap to zero at the rates of Theorems 1–3, so the algorithm reaches an ε-stationary point of Φ that, by gradient alignment, is an ε-approximate Nash equilibrium. Where invoked, Assumption 5 only lets the potential gap $\Phi^\dagger - \Phi$ additionally be read as a *global* optimality gap.

Theorem 1. *Consider Algorithm 1 with biased momentum, step size* $\eta_t = 1/T$*, and* $\rho_t = 4/(t+8)^{2/3}$*. The bound below holds under both feedback models of Definition 3; only the variance constant* σ^2 *differs (the bandit case carries an extra* $\mathcal{O}(\log T)$ *factor, leaving the rate exponent unchanged). Under Assumptions 2–5, define* $\Phi^\dagger = \max_{\pi' \in \Delta} \Phi(\pi')$*. The output* $\bar\pi^T = \frac{1}{T}\sum_{t=1}^T \pi^t$ *satisfies:*

$$\mathbb{E}\big[\Phi^\dagger - \Phi(\bar\pi^T)\big] \le \frac{\Phi^\dagger - \Phi(\pi^1)}{e} + \frac{3D\sqrt{Q}}{T^{1/3}} + \frac{LD^2}{2T} \tag{17}$$

Nash Equilibrium Convergence. We now translate potential convergence to Nash-equilibrium guarantees using the potential game structure.

Theorem 2. *Under the conditions of Theorem 1, and Assumption 5, the expected Nash gap of the time-averaged strategy satisfies (under either feedback model):*

$$\mathbb{E}\big[\mathsf{NG}(\bar\pi^T)\big] \le \frac{\Phi^\dagger - \Phi(\pi^1)}{e} + \frac{3D\sqrt{Q}}{T^{1/3}} + \frac{LD^2}{2T} \tag{18}$$

Corollary 2. *To achieve an* ε*-approximate Nash equilibrium, i.e.,* $\mathbb{E}[\mathsf{NG}(\bar\pi^T)] \le \varepsilon$*, Algorithm 1 with biased momentum requires:*

$$T = \mathcal{O}\left(\frac{D^3 Q^{3/2}}{\varepsilon^3}\right) = \mathcal{O}\left(\frac{n^6}{\varepsilon^3}\right) \text{ rounds.} \tag{19}$$

As shown in Remark 2 and Example 1, Assumption 5 generally fails, so Theorems 1 and 2 are to be read as convergence to a stationary point of Φ at the stated rates, with the Nash gap controlled by the unconditional Corollary 1 rather than by a concavity argument.

Theorem 3. *Consider Algorithm 1 with the unbiased momentum estimator* (12)*. Under Assumptions 2–5, and assuming the stochastic Jacobian-vector product estimator satisfies:*

$$\mathbb{E}\big[\|\widehat{\nabla}_\pi(\nabla_{\pi_i}\Phi)(\pi) - \nabla_\pi(\nabla_{\pi_i}\Phi)(\pi)\|_\Phi^2\big] \le \sigma_H^2 \tag{20}$$

for some $\sigma_H < \infty$*, the time-averaged iterate* $\bar\pi^T$ *satisfies:*

$$\mathbb{E}\big[\mathsf{NG}(\bar\pi^T)\big] \le \frac{\Phi^\dagger - \Phi(\pi^1)}{e} + \mathcal{O}\left(\frac{D(\sigma + \sigma_H D/T)}{\sqrt{T}}\right) \tag{21}$$

Corollary 3. *To achieve* $\mathbb{E}[\mathsf{NG}(\bar\pi^T)] \le \varepsilon$ *(ignoring the initialisation term), Algorithm 1 with unbiased momentum requires:*

$$T = \mathcal{O}\left(\frac{D^2(\sigma^2 + \sigma_H^2)}{\varepsilon^2}\right) = \mathcal{O}\left(\frac{n^4}{\varepsilon^2}\right)$$

rounds, improving the ε*-dependence from cubic to quadratic compared to the biased variant.*

Remark 3. *Both variants achieve sublinear regret:* $\mathcal{R}^T \to 0$ *as* $T \to \infty$*. The unbiased variant achieves the optimal* $\mathcal{O}(1/\sqrt{T})$ *rate for stochastic optimization, while the biased variant trades off a slower* $\mathcal{O}(1/T^{1/3})$ *rate for reduced per-iteration complexity (no Hessian computation required).*

4 EXPERIMENTAL ANALYSIS

We evaluate the proposed algorithms on synthetic instances and a real-world crowdsourcing dataset, assessing Nash-gap convergence, cumulative regret, scalability, and sensitivity to hyperparameters. All results are averaged over 10 independent trials; shaded bands denote ± 1 s.d.[3].

[3]All code, data, and plots are publicly released via GitHub.

Datasets. We use the GMISSION crowdsourcing platform [Chen et al., 2014] (n=20, m=50), with agent preferences constructed from geographic co-location similarity ($\bar{p}_i(k) = \cos(\mathbf{l}_i, \mathbf{l}_k) \in [-1, 1]$, symmetrised). Synthetic instances span three scales: **small** (n=5, m=10, T=300), **medium** (n=20, m=50, T=500), and **large** (n=100, m=200, T=1000), with $r(j) \sim \mathrm{Unif}(1, 10)$ and $\bar{p}_i(k) \sim \mathrm{Unif}(-1, 1)$ (symmetrised), γ=1.

Baselines. We compare `DOS-PGA` (Biased) (Alg. 3, Option A, $\mathcal{O}(T^{-1/3})$) and `DOS-PGA` (Unbiased) (Option B, $\mathcal{O}(T^{-1/2})$) against `UCB-NE` (Alg. 1+2, optimistic oracle), `Naive-SGD` (unbiased single-sample, no momentum), **Uniform Exploration**, and a `Pair-Feature UCB` baseline (the general bandit potential-game solver of Cohen et al. [2017] instantiated with pairwise features), that maintains an independent UCB estimate of every pairwise preference $\bar{p}_i(k)$ and best-responds on the resulting utility estimates. The latter uses the same optimistic feedback as `DOS-PGA` but discards the potential-alignment structure, thereby isolating the contribution of the shared-potential gradient oracle from that of optimism alone; consistent with our $\mathcal{O}(n^4/\varepsilon^2)$ vs. $\mathcal{O}(n^6/\varepsilon^3)$ analysis, it quantifies the n^2 gain from exploiting ASHPG structure.

Figure 1 shows the Nash gap $\mathbb{E}[\widehat{\mathrm{NG}}(\bar{\pi}^T)]$ across the three synthetic scales. Both DOS-PGA variants converge sublinearly and closely track the theoretical $\mathcal{O}(t^{-1/3})$ and $\mathcal{O}(t^{-1/2})$ reference slopes overlaid on each panel. Figure 2 compares DOS-PGA (Biased) under the two feedback models of Definition 3 on the medium synthetic instance (n=15, m=30, T=350). Both variants converge sub-linearly, but the semi-bandit variant achieves a materially lower Nash gap at every horizon (approximately a factor of two at T=350), because disaggregated observations yield tighter preference estimates and less exploration noise. This gap is expected to widen with n, as the bandit estimator cannot disentangle changes in reward-sharing from preference effects alone.

Momentum constant (Figure 3a), social scale γ (Figure 3b), and negative preferences (Figure 3c). Sweeping the momentum constant c in $\rho_t = c/(t+8)^{2/3}$ shows that the theoretically prescribed $c = 4$ achieves the best Nash gap, while smaller values slow early progress by under-weighting fresh gradients and larger values increase variance, hindering later convergence; performance remains robust for $c \in \{2, 4\}$. Increasing the social scale γ raises both the Lipschitz constant and gradient variance, slowing convergence as predicted, yet the Nash gap remains below 0.15 within 350 rounds even for large γ, indicating robustness to strong social preferences, with $\gamma = 0$ yielding the fastest convergence and no qualitative change observed across values. Finally, although our theory assumes non-negative preferences, experiments with mixed-sign preferences show only a modest ($\sim 1.3\times$) slowdown and comparable final Nash gaps, suggesting the assumption is not restrictive in practice.

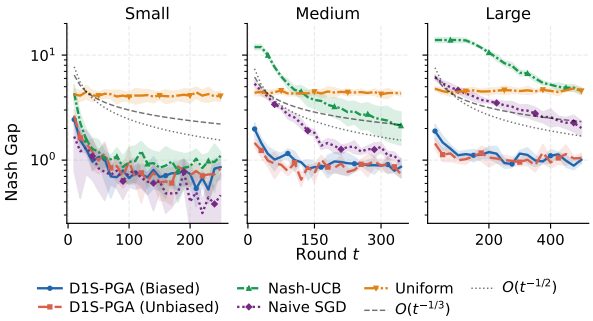

Figure 1: Nash gap vs. round across three synthetic scales.

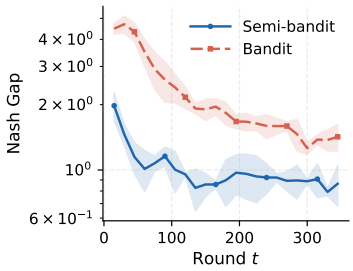

Figure 2: DOS-PGA (Biased) under semi-bandit vs. bandit feedback (n=15, m=30, T=350).

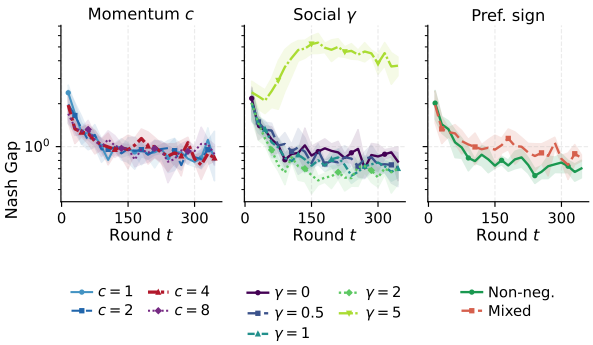

Figure 3: Ablation studies on a 12-agent, 25-project instance (T=350).

5 CONCLUSION

We introduced ASHPGs, a unified model of agents' joint preferences over rewards and coalition composition. Under symmetric preferences, gradient alignment enables a single variance-reduced oracle (DOS-PGA) to achieve both regret minimisation and Nash convergence: the biased SCG variant attains an $\mathcal{O}(T^{-1/3})$ Nash gap with $\mathcal{O}(n^6/\varepsilon^3)$ samples, while the unbiased 1-SFW variant achieves $\mathcal{O}(T^{-1/2})$ and $\mathcal{O}(n^4/\varepsilon^2)$ with stochastic Hessian-vector products. Experiments on synthetic data and GMISSION validate the theory and show the benefit of momentum-based variance reduction. Future work includes non-concave preferences, asynchronous and communication-limited settings, bandit feedback, and dynamic preferences.

References

John Augustine, Ning Chen, Edith Elkind, Angelo Fanelli, Nick Gravin, and Dmitry Shiryaev. Dynamics of profit-sharing games. *Internet Mathematics*, 11(1):1–22, 2015.

Haris Aziz and Florian Brandl. Existence of stability in hedonic coalition formation games. *arXiv preprint arXiv:1201.4754*, 2012.

Haris Aziz and et al. Stable matching under preferences over colleagues. In *IJCAI*, 2019.

Haris Aziz and Rahul Savani. Hedonic games. 2016.

Haris Aziz, Florian Brandl, Felix Brandt, Paul Harrenstein, Martin Olsen, and Dominik Peters. Fractional hedonic games. *ACM Transactions on Economics and Computation (TEAC)*, 7(2):1–29, 2019.

Maria-Florina Balcan, Ariel D Procaccia, and Yair Zick. Learning cooperative games. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 475–481, 2015.

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

Suryapratim Banerjee, Hideo Konishi, and Tayfun Sönmez. Core in a simple coalition formation game. *Social Choice and Welfare*, 18(1):135–153, 2001.

Vittorio Bilò, Angelo Fanelli, Michele Flammini, Gianpiero Monaco, and Luca Moscardelli. Nash stable outcomes in fractional hedonic games: Existence, efficiency and computation. *Journal of Artificial Intelligence Research*, 62:315–371, 2018.

Vittorio Bilò, Laurent Gourvès, and Jérôme Monnot. Project games. *Theoretical Computer Science*, 940:97–111, 2023.

Niclas Boehmer, Martin Bullinger, and Anna Maria Kerkmann. Causes of stability in dynamic coalition formation. *ACM Transactions on Economics and Computation*, 2023.

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

Felix Brandt, Martin Bullinger, and Anaëlle Wilczynski. Reaching individually stable coalition structures in hedonic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5211–5218, 2021.

Martin Bullinger and René Romen. Stability in online coalition formation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9537–9545, 2024.

Katarı´ na Cechlárová and Antonio Romero-Medina. Stability in coalition formation games. *International Journal of Game Theory*, 29:487–494, 2001.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Georgios Chalkiadakis, Nicholas R Jennings, and Craig Boutilier. Learning, coalition formation and coordination. In *AAAI*, 2008.

Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational and gametheoretic aspects of coalition formation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011.

Zhao Chen, Rui Fu, Ziyuan Zhao, Zheng Liu, Leihao Xia, Lei Chen, Peng Cheng, Caleb Chen Cao, Yongxin Tong, and Chen Jason Zhang. gmission: A general spatial crowdsourcing platform. *Proceedings of the VLDB Endowment*, 7(13):1629–1632, 2014.

Gabriel Cohen and Noa Agmon. Online learning of stable partitions in hedonic games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1234–1240, 2024.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Saar Cohen and Noa Agmon. Decentralized online learning by selfish agents in coalition formation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025a.

Saar Cohen and Noa Agmon. Online learning of coalition structures by selfish agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13709–13717, 2025b.

Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. Learning in congestion games with bandit feedback. *Advances in Neural Information Processing Systems*, 35: 11009–11022, 2022.

Andreas Darmann. Group activity selection from ordinal preferences. In *International Conference on Algorithmic Decision Theory*, pages 35–51. Springer, 2015.

Andreas Darmann, Edith Elkind, Sascha Kurz, Jérôme Lang, Joachim Schauer, and Gerhard Woeginger. Group activity selection problem. In *Internet and Network Economics: 8th International Workshop, WINE 2012, Liverpool, UK, December 10-12, 2012. Proceedings 8*, pages 156–169. Springer, 2012.

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

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

Edith Elkind and Michael J Wooldridge. Hedonic coalition nets. In *AAMAS (1)*, pages 417–424. Citeseer, 2009.

Simone Fioravanti, Michele Flammini, Bojana Kodric, and Giovanna Varricchio. Pac learning and stabilizing hedonic games: towards a unifying approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5641–5648, 2023.

Michel Goemans, Li Erran Li, Vahab S Mirrokni, and Marina Thottan. Market sharing games applied to content distribution in ad-hoc networks. In *Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing*, pages 55–66, 2004.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Hooyeon Lee and Virginia Vassilevska Williams. Parameterized complexity of group activity selection. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 353–361, 2017.

Haipeng Liu et al. Sharp regret bounds for online learning in multi-agent games. *Journal of Machine Learning Research*, 22:1–40, 2021a.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021b.

Jason R Marden and Tim Roughgarden. Generalized efficiency bounds in distributed resource allocation. In *49th IEEE Conference on Decision and Control (CDC)*, pages 2233–2238. IEEE, 2010.

Jason R Marden and Adam Wierman. Distributed welfare games. *Operations Research*, 61(1):155–168, 2013.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *International Conference on Artificial Intelligence and Statistics*, pages 1886–1895. PMLR, 2018.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 21(105):1–49, 2020.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

Jakub Sliwinski and Yair Zick. Learning hedonic games. In *IJCAI*, pages 2730–2736, 2017.

Jaber Valizadeh and Ray Telikani. Online learning for decentralized multi-agent planning in repeated hedonic skill games. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 36, pages 342–350, 2026.

Jaber Valizadeh, Dongmo Zhang, and Omar Mubin. Autonomy with structural task allocation games: From inefficiency to optimality. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 435–452. Springer, 2025a.

Jaber Valizadeh, Dongmo Zhang, and Omar Mubin. Learning preferences in additive separable hedonic project games. In *Australasian Joint Conference on Artificial Intelligence*, pages 491–505. Springer, 2025b.

Adrian Vetta. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 416–425. IEEE, 2002.

Berthold Vöcking. Selfish load balancing. *Algorithmic game theory*, 20:517–542, 2007.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.