Offline Learning of Nash Stable Coalition Structures with Possibly Overlapping Coalitions

AAAI Track

Saar Cohen
Department of Computer Science, Bar Ilan University
Ramat Gan, Israel
Department of Computer Science, University of Oxford
Oxford, United Kingdom
saar30@gmail.com

ABSTRACT

Coalition formation concerns strategic collaborations of selfish agents that form coalitions based on their preferences. It is often assumed that coalitions are disjoint and preferences are fully known, which may not hold in practice. In this paper, we thus present a new model of coalition formation with *possibly overlapping* coalitions under *partial information*, where selfish agents may be part of *multiple* coalitions simultaneously and their full preferences are initially unknown. Instead, information about past interactions and associated utility feedbacks is stored in a fixed offline dataset, from which we aim to efficiently infer agents' preferences. We analyze the impact of diverse dataset information constraints by studying two utility feedback models: *semi-bandit* (agent-level) and *bandit* (coalition-level) feedbacks. For both models, we identify assumptions under which the dataset covers sufficient information for an offline learning algorithm to infer preferences and use them to recover a partition that is (approximately) *Nash stable*, i.e., no agent can improve her utility by unilaterally deviating. We also aim to devise algorithms with *low sample complexity*, requiring only a small dataset to obtain a desired approximation to Nash stability. Under semi-bandit feedback, we provide a sample-efficient algorithm proven to obtain an approximately Nash stable partition under a *sufficient* and *necessary* assumption on the information covered by the dataset. Yet, under bandit feedback, we show that only a stricter assumption is sufficient for sample-efficient learning. Still, in multiple cases, our algorithms' sample complexity bounds have ***optimality*** guarantees up to logarithmic factors. Finally, extensive experiments show our algorithm's approximation to Nash stability.

KEYWORDS

Coalition Formation; Nash Stability; Offline Learning

ACM Reference Format:
Saar Cohen. 2026. Offline Learning of Nash Stable Coalition Structures with Possibly Overlapping Coalitions: AAAI Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/ZZZS1209

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ZZZS1209

1 INTRODUCTION

In a large consulting firm, managers must assign consultants to upcoming client projects across various domains (e.g., finance, logistics). A consultant may work on multiple projects, one per domain. The compatibility between consultants depends on the domain, e.g., two employees may collaborate efficiently on logistics, but clash in finance due to differing approaches. However, the managers do not have full access to employees' preferences over team compositions. Testing new team structures through trial projects is also infeasible due to the financial and organizational costs of reassignments, client-facing risks, and time constraints. Instead, managers rely on a fixed dataset of historical project outcomes and post-project peer evaluations. Based only on this limited, pre-collected information, they aim to assign consultants to teams that align with their true but unknown preferences, ensuring no one would prefer to switch groups unilaterally. Such scenarios and many other real-world cases exemplify *coalition formation*, where *agents* perform activities in *coalitions* rather than on their own, and the challenge is forming coalitions that satisfy some desired criteria.

A popular model for studying coalition formation is that of *hedonic games* [24], whose outcome is a set of disjoint coalitions (hereafter, *partition*). The desirability of partitions is often evaluated in terms of various concepts of *single-agent stability* based on agents' preferences [2, 9], where no agent benefits from leaving her current coalition to join another one on her own. A common such notion is *Nash stability*, where no agent can increase her utility by unilaterally deviating (e.g., no consultant can improve by changing teams unilaterally). In the hedonic games literature, agents express preferences only for coalitions they are part of while disregarding inter-coalitional relationships, and these preferences are typically assumed to be fully known when computing stable partitions. Yet, both assumptions may not hold in many real-life scenarios, as in our consulting firm example.

In this paper, we thus introduce and study a *new* framework of coalition formation with *possibly overlapping* coalitions in *partial information* settings, where selfish agents may join *several* coalitions concurrently, following either *mixed* or *pure* strategies, while their full preferences are initially unknown. Instead, we only have partial knowledge about past interactions and corresponding utility feedbacks, which is stored in a fixed offline dataset. The dataset is collected in advance, without further interactions within the environment, reflecting real-life settings where active interactions may be costly and risky, like in our consulting firm example. Our goal is

to maximally exploit the available dataset to efficiently infer agents' preferences. We exhibit our results for *additively separable* and *symmetric* preferences [8], where any pair of agents assigns the same cardinal utility to one another, and an agent's utility for her chosen coalitions is the *sum* of her utilities from their members. However, agents' mutual utilities for one another may vary across different coalitions, reflecting real-world cases such as our consulting firm example, where a consultant may benefit more from another peer in one domain than in another. Symmetric preferences are realistic in reciprocal interactions like friendships, as widely studied in friend-oriented hedonic games [23, 28]. Further, our results for *mixed* strategies readily extend to *asymmetric* preferences, where agents may value each other differently, but extending our analysis for *pure* strategies may be generally infeasible (see Remark 2).

Many realistic scenarios differ in the type and granularity of utility information available in the dataset. We thus explore the effect of different information constraints by studying two forms of utility feedbacks that can be stored in the dataset. Specifically, we examine *semi-bandit* utility feedbacks, where *agent-level* utilities are available for each agent's interactions with other members of her chosen coalitions. We also analyze *bandit* feedback settings with less granular utility information, where we can only observe *coalition-level* utility feedbacks about each agent's overall utility from her chosen coalitions.

Contributions. In both semi-bandit and bandit feedback settings, we characterize assumptions under which the dataset covers sufficient information for an offline learning algorithm to deduce agents' preferences and use them to construct an outcome that is (approximately) Nash stable. Under those assumptions, we develop algorithms with *low sample complexity*, requiring only a small dataset to reach a desired approximation to Nash stability. In semi-bandit feedback settings, we design a sample-efficient offline learning algorithm that attains an approximately Nash stable outcome under a minimal assumption on the information covered by the dataset. Our assumption is *sufficient* and *necessary*, as we prove that no weaker assumption enables efficient learning of an approximately Nash stable outcome, regardless of the dataset size. Intuitively, we require that the dataset reflects the partitions that may form through unilateral deviations. For instance, in our consulting firm example, if a unilateral deviation would result in a finance team of a certain size, then the dataset should include some past project outcomes with a finance team of that same size. However, under bandit feedback, we show that only a stricter assumption is *sufficient* for sample-efficient learning of an approximately Nash stable outcome. Specifically, for any agent and any unilateral deviation of that agent from some Nash stable outcome, the stricter assumption requires that the dataset is at least as informative as a sufficiently large dataset generated according to the outcome of that deviation. In many cases, our algorithms' sample complexity bounds have ***optimality*** guarantees up to logarithmic factors. Finally, extensive experiments confirm that our algorithms consistently reaches a low approximation to Nash stability in a variety of settings.

2 RELATED WORKS

Hedonic games have been presented by Dréze and Greenberg [24], and later expanded to the study of various notions of stability, fairness, and optimality (see, e.g., [2, 44]). Highly related to our work are *additively separable hedonic games* (ASHGs) with symmetric preferences [8], where many works evaluate the system by means of Nash stability [1, 5, 6]. Particularly, Bogomolnaia and Jackson [8] proved that Nash stable partitions may not exist in **general** ASHGs, while Sung and Dimitrov [40] showed that checking if an instance admits such a partition is NP-complete in the strong sense. In contrast, ASHGs with ***symmetric*** preferences admit a Nash stable partition due to a potential function argument [8], but computing such partitions is PLS-complete [28]. However, while hedonic games do not allow agents to be part of several coalitions, our model captures realistic cases where agents can join *multiple* coalitions, which thus may overlap.

Shehory and Kraus [35, 36] introduced the first model of overlapping coalition formation for handling task allocation, which was later followed by additional works on task-oriented applications (see, e.g., [22, 32, 47, 48]). However, this line of research aims to find (approximately) optimal coalition structures, while we analyze the system by means of stability. Though the *group* stability notion of the core has been examined in cooperative games with overlapping coalitions (see, e.g., [10, 25, 52, 53]), we explore Nash stability, which is based on *single-agent* deviations. Further, all above works on overlapping coalition formation typically allow arbitrary monetary transfers, i.e., the payoff or resources of a coalition can be distributed arbitrarily among its members. Conversely, monetary transfers are *unavailable* in our context, as we consider games with *non-transferable* utilities.

However, the above works on hedonic games and cooperative games with overlapping coalitions unrealistically require that the agents' preferences are fully known when computing stable partitions. To tackle this issue, several studies have examined PAC learning in those domains [4, 26, 30, 39], using samples to learn agents' preferences and a *core-stable* partition, where no subset of agents can improve their utility by regrouping into a new coalition. Yet, while they focus on the *group* deviations, we study *single-agent* deviations by analyzing the system in terms of Nash stability. Recently, Cohen and Agmon [14, 15, 16, 17, 18, 19] and Cohen [11] proposed *online* and *online learning* variants of coalition formation. In online learning of coalition structures, agents initially lack preference knowledge and form coalitions based on preferences learned through active interactions. While Cohen and Agmon [17, 19] analyze Nash stability in *online* settings, we assess Nash stability in *offline* scenarios where active interactions are costly or risky. Instead, agents' preferences are inferred from a fixed, pre-collected dataset without further interactions.

Our work is also closely related to offline reinforcement learning (RL), aiming to learn an optimal policy from a dataset collected a priori without further interactions with the environment. A key challenge in offline RL is the insufficient coverage of the dataset [43], arising from the lack of continued exploration [41]. To address this challenge, existing studies presented various assumptions on the sufficient coverage of the dataset in both single-agent settings [31, 34, 42, 46] and multi-agent settings [38, 49, 50]. Recently, minimal assumptions for offline zero- and general-sum games have been identified [20, 21, 51], together with algorithms for learning a Nash equilibrium. However, their sample complexity scales with the number of actions each agent can take, which may be *exponential*

in the number of agents in our setting. In contrast, our proposed assumptions allow for algorithms that remove this exponential dependency, while attaining sample complexity with ***optimality*** guarantees (up to logarithmic factors).

3 PRELIMINARIES

We consider a ***possibly overlapping coalition formation*** (POCF) game $G = (\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}})$, defined as follows. We are given a finite set $\mathcal{N} = \{1, \dots, n\}$ of n selfish agents with *unknown* preferences. Hereafter, we denote $[k] := \{1, \dots, k\}$ for $k \in \mathbb{N}$ and $[0] = \{0\}$. We focus on realistic scenarios where the number of coalitions may be constrained. For an integer $k \geq 1$, each agent can join one of k *candidate* coalitions. In our consulting firm example, this can be thought of as if each consultant picks which rooms to enter among k rooms. Thus, in our setting of possibly overlapping coalitions, the ***action space*** of each agent i is denoted by $\mathcal{A}_i \subseteq \mathcal{P}([k]) \setminus \{\emptyset\}$, where $\mathcal{P}([k])$ is the power set of $[k]$ and each action $a_i \in \mathcal{A}_i$ is a subset of at most k candidate coalitions that agent i can join, where $a_i \neq \emptyset$ (i.e., agent i always decides to join some coalition). Note that, if $|a_i| \leq 1$ for any agent i and any action $a_i \in \mathcal{A}_i$, then each agent can participate in at most one coalition, in which case coalitions are *disjoint*; otherwise, coalitions may *overlap*.

Each agent i can join certain coalitions among k candidate ones following a ***mixed strategy*** $\varphi_i \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the probability simplex over $\mathcal{A}_i$, i.e., for any $a_i \in \mathcal{A}_i$, agent i joins the candidate coalitions in a_i with probability $\varphi_i(a_i) \in [0, 1]$. Letting $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ be the *joint action space*, each agent i can then sample an action $a_i \in \mathcal{A}_i$ from φ_i independently from other agents, forming a *joint action* $\mathbf{a} = (a_i)_{i\in\mathcal{N}}$. Similarly, let $\boldsymbol{\varphi} = (\varphi_i)_{i\in\mathcal{N}}$ be the agents' *joint mixed strategy*. The constructed joint action $\mathbf{a}$ induces a partition of the agents $\pi^{\mathbf{a}} = (C_\ell^{\mathbf{a}})_{\ell\in[k]}$ with possibly overlapping coalitions, where, for any $\ell \in [k]$, $C_\ell^{\mathbf{a}}$ is the set of agents joining the ℓth candidate coalition, i.e., $C_\ell^{\mathbf{a}} = \{i \in \mathcal{N} : \ell \in a_i\}$. For any $\ell \in [k]$, if no agent joins the ℓth candidate coalition (i.e., $\ell \notin a_i$ for any agent i), then this coalition is empty. Thereby, we denote by $|\pi^{\mathbf{a}}|$ the number of non-empty coalitions in $\pi^{\mathbf{a}}$. We also denote the coalitions in $\pi^{\mathbf{a}}$ containing agent i as $\pi^{\mathbf{a}}(i)$.

Afterwards, we can derive each agent's utility from her chosen strategy, determined by aggregating her utilities resulting from interactions with other agents. We focus on POCF games with *additively separable* and *symmetric*[1] valuations in each coalition, where the utility that any pair of agents derives from their interaction within that coalition is equal, indicating the intensity by which they prefer each other to another agent. However, their mutual utilities from each other may differ across different coalitions. This captures real-life scenarios, such as our consulting firm example, where a consultant may benefit more from another one in one domain than in another. An agent's utility from her chosen coalitions is then the sum of her utilities from other members of those coalitions. As common in the literature (see, e.g., [27]), we assume that agents' utilities are within $[-1, 1]$. Recall that preferences are *unknown* and even the agents themselves may not be aware of them. Thus, supposing a joint action $\mathbf{a}$ is chosen by all agents, the uncertainty about the mutual valuations resulting from their interactions within the ℓ-th candidate coalition is captured by an *unknown* and *fixed* distribution $\mathcal{D}_{i,j}^{\ell}(\cdot|\mathbf{a})$ over $[-1, 1]$ with mean $d_{i,j}^{\ell}$ for any pair of distinct agents i, j, which we aim to learn. The ***mutual utility*** $v_{i,j}^{\ell}$ of agents i, j that results from their interactions in the ℓ-th candidate coalition according to a joint action $\mathbf{a}$ is then independently drawn from $\mathcal{D}_{i,j}^{\ell}(\cdot|\mathbf{a})$. We use the convention that $v_{i,i}^{\ell} = d_{i,i}^{\ell} = 0$ for any agent i. Thus, for any joint action $\mathbf{a}$ sampled from a joint mixed strategy $\boldsymbol{\varphi}$, if agent i decides to join some candidate coalitions in a_i, then her utility from the induced partition $\pi^{\mathbf{a}} = (C_\ell^{\mathbf{a}})_{\ell\in[k]}$ is $v_i(\mathbf{a}) = \sum_{\ell\in a_i} \sum_{i\neq j\in C_\ell^{\mathbf{a}}} v_{i,j}^{\ell}$, whose mean is $d_i(\mathbf{a}) = \sum_{\ell\in a_i} \sum_{i\neq j\in C_\ell^{\mathbf{a}}} d_{i,j}^{\ell}$. Agent i's utility from her strategy φ_i is thus defined as $V_i(\boldsymbol{\varphi}) := \mathbb{E}_{\mathbf{a}\sim\boldsymbol{\varphi}}[d_i(\mathbf{a})]$.

3.1 Nash Stability in POCF Games

Each agent joins a coalition with the goal of maximizing her own utility. We thus want to study stability und er single agents' incentives to deviate between coalitions. When agents play ***pure*** strategies, each agent i's strategy is joining certain candidate coalitions by only picking some action $a_i \in \mathcal{A}_i$. Let $\mathbf{a}_{-i} = (a_j)_{j\neq i}$ be the joint strategy of all agents except for agent i. Agent i can then ***deviate*** by moving from her selected coalition to another one with index $a_i' \in \mathcal{A}_i$, which is a ***Nash deviation*** if it improves her utility, i.e., $V_i(\mathbf{a}_{-i}, a_i') > V_i(\mathbf{a}_{-i}, a_i)$. We also study *mixed* strategies, where we consider the other notion of *mixed Nash deviations*. Consider the joint strategy $\boldsymbol{\varphi}$. Letting $\boldsymbol{\varphi}_{-i} = (\varphi_j)_{j\neq i}$ for any agent i, agent i may perform a ***(mixed single-agent) deviation*** from her strategy φ_i to another strategy $\phi_i \in \Delta(\mathcal{A}_i)$, which is a ***mixed Nash deviation*** only if it immediately makes her better off, i.e., $V_i(\boldsymbol{\varphi}_{-i}, \phi_i) > V_i(\boldsymbol{\varphi})$. Hence, a (*pure* or *mixed*) joint strategy for which no Nash deviation is possible is said to be ***Nash stable*** (NS), or a *Nash equilibrium*.

Our POCF class of coalition formation games generalizes the well-known class of *symmetric additively separable hedonic games* (S-ASHGs) [8]. Indeed, consider cases where the number of coalitions is unconstrained ($k = n$), and each agent can join exactly one of the n *candidate* coalitions (i.e., $\mathcal{A}_i = \{\{\ell\}\}_{\ell\in[n]}$ for any agent i). We thereby obtain S-ASHGs. For S-ASHGs under *pure* strategies, the existence of a Nash stable outcome is guaranteed by potential function argument [8]. Next, we show that our class of symmetric POCF games are also potential games, establishing that they always admit at least one pure (and hence mixed) NS strategy. Pure NS strategies need not be unique (e.g., if all agents assign zero utility to one another, then any pure strategy is Nash stable).

Lemma 1. *A symmetric POCF game with unknown, symmetric preferences is a* potential game. *Therefore, any symmetric POCF game always admits at least one pure (and thus mixed) NS strategy.*

Proof. (*Sketch*) Consider a joint mixed strategy $\boldsymbol{\varphi}$. Given a joint action $\mathbf{a} \sim \boldsymbol{\varphi}$, in [12, Appendix A] we prove that $\Phi(\mathbf{a}) = \frac{1}{2}\sum_{i\in\mathcal{N}} v_i(\mathbf{a})$ is a potential function for *pure* strategies, as $\Phi(\mathbf{a}_{-i}, a_i) - \Phi(\mathbf{a}_{-i}, a_i') = v_i(\mathbf{a}_{-i}, a_i) - v_i(\mathbf{a}_{-i}, a_i')$ for any agent i playing some action $a_i' \neq a_i$. Similarly, slightly abusing notation, $\Phi(\boldsymbol{\varphi}) = \frac{1}{2}\sum_{i\in\mathcal{N}} V_i(\boldsymbol{\varphi})$ is a potential for *mixed* strategies as $\Phi(\boldsymbol{\varphi}_{-i}, \varphi_i) - \Phi(\boldsymbol{\varphi}_{-i}, \phi_i) = V_i(\boldsymbol{\varphi}_{-i}, \varphi_i) - V_i(\boldsymbol{\varphi}_{-i}, \phi_i)$ for any agent i playing another strategy $\phi_i \in \Delta(\mathcal{A}_i)$. □

[1] As noted in Section 1, our results for *mixed* strategies naturally extend to *asymmetric* preferences, yet this is generally infeasible for *pure* strategies (see Remark 2).

In S-ASHGs with *pure* strategies, computing NS strategies is PLS-complete [28]. As S-ASHGs are a special case of our model, this PLS-completeness extends to symmetric POCF games by Lemma 1. We thus also consider an *approximate* notion of Nash stability. For a joint strategy $\boldsymbol{\varphi}$, agent i's *best response* to the other agents' strategies is a Nash deviation given by a strategy $\phi_i^\star$ satisfying $V_i^\star(\boldsymbol{\varphi}_{-i}) := V_i(\boldsymbol{\varphi}_{-i}, \phi_i^\star) = \max_{\phi \in \Delta(\mathcal{A}_i)} V_i(\boldsymbol{\varphi}_{-i}, \phi)$. We measure the worst agent's local gap between the expected utilities she gets from her best response and her current strategy by $\boldsymbol{\varphi}$'s ***duality gap***:

$$\mathrm{Gap}(\boldsymbol{\varphi}) := \max_{i \in \mathcal{N}} [V_i^\star(\boldsymbol{\varphi}_{-i}) - V_i(\boldsymbol{\varphi})] \tag{1}$$

Hence, for any $\varepsilon > 0$, the agents' joint strategy $\boldsymbol{\varphi}$ is ε-***approximate Nash stable*** (ε-NS) if no agent can improve her gain by more than ε, i.e., $\mathrm{Gap}(\boldsymbol{\varphi}) \leq \varepsilon$. If $\mathrm{Gap}(\boldsymbol{\varphi}) = 0$, then $\boldsymbol{\varphi}$ is *exactly* Nash stable.

Remark 1. *If $\{\ell\} \in A_i$ for any $\ell \in [k]$ and any agent i, consider the mixed strategy φ_i where each agent treats all candidate coalitions as equally desirable, i.e., $\varphi_i(\{\ell\}) = \frac{1}{k}$ for any $\ell \in [k]$. Clearly, this is an exact mixed NS strategy, but it ignores the agents' preferences entirely, which is unrealistic in practical scenarios as agents often act strategically based on their own preferences, not arbitrarily. Instead, our framework aims to learn preference-aligned mixed NS strategies that reflect real-life behavior by accounting for agents' incentives.*

3.2 Offline Learning in Coalition Formation

In the offline version of our model, we have access only to a fixed dataset $\mathcal{S} = \{(\mathbf{a}^m, \mathbf{v}^m)\}_{m=1}^M$ of M samples independently[2] drawn from a possibly unknown exploration policy $\rho \in \Delta(\mathcal{A})$, where $\mathbf{a}^m$ is a joint action sampled from ρ and $\mathbf{v}^m$ comprises utility feedbacks resulting from the partition induced by $\mathbf{a}^m$. In particular, no further sampling from ρ is allowed. We say that a joint action $\mathbf{a}$ is ***covered*** by the exploration policy if $\rho(\mathbf{a}) > 0$. We study the following two utility feedback models, ordered by decreasing level of granularity:

(1) **Semi-bandit feedback,** capturing scenarios where we can observe *agent-level* feedback about each agent's utilities from her interactions with any other member of her chosen coalitions, i.e., for any $m \in [M]$, each agent i and any $\ell \in a_i^m$, we attain her realized utility $v_{i,j}^{\ell,m}$ for each agent $j \neq i$ in the ℓ-th candidate coalition $C_\ell^{\mathbf{a}^m}$ formed by $\mathbf{a}^m$, yielding $\mathbf{v}^m = \{v_{i,j}^{\ell,m}\}_{i \in \mathcal{N}, \ell \in a_i^m, i \neq j \in C_\ell^{\mathbf{a}^m}}$.

(2) **Bandit feedback,** reflecting cases where we can only observe *coalition-level* feedback about each agent's overall utility from her coalitions. Namely, for any $m \in [M]$ and each agent i, we only obtain agent i's utility from the partition induced by $\mathbf{a}^m$ (i.e., $v_i(\mathbf{a}^m)$), with no information about the individual utility $v_{i,j}^{\ell,m}$ assigned to any agent $j \neq i$, i.e., $\mathbf{v}^m = \{v_i(\mathbf{a}^m)\}_{i \in \mathcal{N}}$.

Under each feedback model, our goal is identifying conditions under which the dataset covers sufficient information for an *offline learning* algorithm to learn preferences and use them to recover an ε-NS joint strategy that aligns with the true preferences. We term such assumptions as *(dataset) coverage assumptions*. Particularly, our goal is devising algorithms with ***small duality gap*** and ***low sample complexity***, i.e., such algorithms find an ε-NS strategy for a small $\varepsilon > 0$ using a number of samples that is small in its dependency on the number of agents n and $1/\varepsilon$.

[2]Independence is a standard assumption in offline learning (e.g., [20, 21]), where the goal is to learn from a fixed dataset rather than model its temporal generation. In our setting, this is also practical: for statistical analysis, data can be treated as i.i.d. samples from an exploration policy, even if it was generated sequentially. Standard subsampling methods (e.g., [20]) can alleviate temporal correlations to enforce independence.

Algorithm 1 Surrogate Minimization in POCF Games

Input: An offline dataset $\mathcal{S} = \{(\mathbf{a}^m, \mathbf{v}^m)\}_{m=1}^M$.

1: Construct $\hat{v}_i$ and b_i^δ for any agent i based on $\mathcal{S}$.

2: Return an *approximate* solution $\boldsymbol{\varphi}^{\text{out}}$ to:

$$\min_{\boldsymbol{\varphi} \in \prod_{i=1}^n \Delta(\mathcal{A}_i)} \max_{i \in \mathcal{N}} \left[\overline{V}_i^{\star,\delta}(\boldsymbol{\varphi}_{-i}) - \underline{V}_i^\delta(\boldsymbol{\varphi}) \right] \tag{6}$$

where $\overline{V}_i(\boldsymbol{\varphi}), \underline{V}_i^\delta(\boldsymbol{\varphi}), \overline{V}_i^{\star,\delta}(\boldsymbol{\varphi}_{-i})$ are as in (3)-(4).

3.3 Surrogate Minimization in POCF Games

In Algorithm 1, we present a general algorithmic framework for learning NS strategies in offline symmetric POCF games that will be used throughout our work. To simplify the discussion, we mainly focus on learning NS *mixed* strategies. Later, in Section 4.2.2 and Footnote 5, we explain how Algorithm 1 can be adapted for learning NS *pure* strategies. Algorithm 1 relies on a ***utility estimator*** $\hat{v}_i : \mathcal{A} \to \mathbb{R}$ for each agent i, which estimates her mean utility from the partition induced by a sampled joint action $\mathbf{a} \in \mathcal{A}$ via $\hat{v}_i(\mathbf{a})$ (line 1). It also exploits an exploration bonus, introducing conservatism into the learning process. Formally, for any agent i and some confidence level $\delta \in (0, 1]$ capturing the degree of certainty, $b_i^\delta : \mathcal{A} \to \mathbb{R}$ is an ***exploration bonus*** for the utility estimator $\hat{v}_i$ if the following holds with probability at least $1 - \delta$ for any joint action $\mathbf{a} \in \mathcal{A}$:

$$|d_i(\mathbf{a}) - \hat{v}_i(\mathbf{a})| \leq b_i(\mathbf{a}) \tag{2}$$

The specific expression for $\hat{v}_i$ and b_i^δ may vary based on the utility feedback model, as we will discuss in the remainder of the paper. In any case, we can define the upper confidence bound (UCB) and the lower confidence bound (LCB) of agent i by $\overline{v}_i^\delta(\mathbf{a}) = \hat{v}_i(\mathbf{a}) + b_i^\delta(\mathbf{a})$ and $\underline{v}_i^\delta(\mathbf{a}) = \hat{v}_i(\mathbf{a}) - b_i^\delta(\mathbf{a})$, respectively. They allow us to construct optimistic and pessimistic estimates of the expected utility obtained by each agent i from a joint strategy $\boldsymbol{\varphi}$, given by (respectively):

$$\overline{V}_i^\delta(\boldsymbol{\varphi}) := \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}}[\overline{v}_i^\delta(\mathbf{a})] \quad , \quad \underline{V}_i^\delta(\boldsymbol{\varphi}) := \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}}[\underline{v}_i^\delta(\mathbf{a})] \tag{3}$$

For a joint strategy $\boldsymbol{\varphi}$, agent i's *optimistic best response* to others' strategies is a Nash deviation given by a strategy $\phi_i^\star$ that satisfies:

$$\overline{V}_i^{\star,\delta}(\boldsymbol{\varphi}_{-i}) := \overline{V}_i^\delta(\boldsymbol{\varphi}_{-i}, \phi_i^\star) = \max_{\phi \in \Delta(\mathcal{A}_i)} \overline{V}_i^\delta(\boldsymbol{\varphi}_{-i}, \phi) \tag{4}$$

Algorithm 1 uses (4) to estimate the true duality gap in (1) via $\widehat{\mathrm{Gap}}^\delta(\boldsymbol{\varphi}) := \max_{i \in \mathcal{N}} [\overline{V}_i^{\star,\delta}(\boldsymbol{\varphi}_{-i}) - \underline{V}_i^\delta(\boldsymbol{\varphi})]$, for which it then computes an approximate minimizer (line 2). Namely, Algorithm 1 finds a joint mixed strategy $\boldsymbol{\varphi}^{\text{out}}$ that solves (6) up to ϵ_{opt}-optimality (e.g., by standard coordinate-descent schemes, as detailed in Section 6)[3]:

$$\widehat{\mathrm{Gap}}^\delta(\boldsymbol{\varphi}^{\text{out}}) \leq \min_{\boldsymbol{\varphi}} \widehat{\mathrm{Gap}}^\delta(\boldsymbol{\varphi}) + \epsilon_{\text{opt}} \tag{5}$$

The intuition behind considering the above estimates is that they serve as surrogates for the true duality gap. Formally:

[3]Solving (6) exactly is generally infeasible, as computing NS outcomes is PLS-complete, even under full information [28].

Lemma 2. *For any $\delta \in (0,1]$ and any joint strategy $\boldsymbol{\varphi}$:* $\mathrm{Gap}(\boldsymbol{\varphi}) \leq \widehat{\mathrm{Gap}}^{\delta}(\boldsymbol{\varphi})$ *and* $\mathrm{Gap}(\varphi^{\mathrm{out}}) \leq \min_{\boldsymbol{\varphi}} \widehat{\mathrm{Gap}}^{\delta}(\boldsymbol{\varphi}) + \epsilon_{\mathrm{opt}}$ *with probability at least $1-\delta$, where φ^{out} is the joint strategy produced by Algorithm 1.*

Proof. (*Sketch*) By (2) and (3), $\underline{V}_i^{\delta}(\boldsymbol{\varphi}) \leq V_i(\boldsymbol{\varphi}) \leq \overline{V}_i^{\delta}(\boldsymbol{\varphi})$ holds with probability at least $1-\delta$. In [12, Appendix B], we show that this easily implies the desired due to (1). □

Next, we supply a general upper bound on the duality gap of the joint strategy φ^{out} produced by Algorithm 1, which serves as a key tool in deriving our main results.

Theorem 1. *Let Γ be the set of all* deterministic *joint strategies, b_i^{δ} be an exploration bonus for the utility estimator $\hat{v}_i$ of any agent i for any $\delta \in (0,1]$ and consider some NS (possibly mixed) joint strategy $\boldsymbol{\varphi}^{\star}$. Then, under each feedback model, the duality gap of the joint strategy φ^{out} produced by Algorithm 1 is upper bounded as follows with probability at least $1-\delta$ (which directly translates into Algorithm 1's approximation guarantees for Nash stability):*

$$\mathrm{Gap}(\varphi^{\mathrm{out}}) \leq 2 \max_{i \in \mathcal{N}} \left[\max_{\boldsymbol{\varphi}' \in \Gamma} \mathbb{E}_{\mathbf{a} \sim (\boldsymbol{\varphi}^{\star}_{-i}, \varphi'_i)} [b_i^{\delta}(\mathbf{a})] + \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}^{\star}} [b_i^{\delta}(\mathbf{a})] \right] + \epsilon_{\mathrm{opt}}$$

Proof. (*Sketch*) In [12, Appendix C], we first prove that (2)-(3) imply that $V_i(\boldsymbol{\varphi}) - \underline{V}_i^{\delta}(\boldsymbol{\varphi})$ and $\overline{V}_i^{\delta}(\boldsymbol{\varphi}) - V_i(\boldsymbol{\varphi})$ are at most $2\mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}}[b_i^{\delta}(\mathbf{a})]$. As $\overline{V}_i^{\delta}(\boldsymbol{\varphi}_{-i}, \phi_i)$ and $\underline{V}_i^{\delta}(\boldsymbol{\varphi})$ are both linear in each entry of $\boldsymbol{\varphi}$ and $\phi_i \in \Delta(\mathcal{A}_i)$ for any agent i, the maximum of $\max_{i \in \mathcal{N}} \left[\overline{V}_i^{\delta}(\boldsymbol{\varphi}_{-i}, \phi_i) - \underline{V}_i^{\delta}(\boldsymbol{\varphi}) \right]$ over all joint strategies $\boldsymbol{\phi} \in \prod_{i=1}^{n} \Delta(\mathcal{A}_i)$ is obtained at a vertex, i.e., a *pure* joint strategy. Combining the above with (1), Lemma 2 and the fact that $\boldsymbol{\varphi}^{\star}$ is an NS (possibly mixed) joint strategy: $\mathrm{Gap}(\varphi^{\mathrm{out}}) \leq \mathrm{Gap}(\boldsymbol{\varphi}^{\star}) + 2\max_{i \in \mathcal{N}} [\max_{\boldsymbol{\varphi}' \in \Gamma} \mathbb{E}_{\mathbf{a} \sim (\boldsymbol{\varphi}^{\star}_{-i}, \varphi'_i)} [b_i^{\delta}(\mathbf{a})] + \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}^{\star}} [b_i^{\delta}(\mathbf{a})]] + \epsilon_{\mathrm{opt}}$. We then obtain the desired from $\mathrm{Gap}(\boldsymbol{\varphi}^{\star}) = 0$. □

Remark 2 (**Asymmetric Preferences**). *Though we focus on symmetric preferences mainly to reduce the valuation space and simplify exposition, our results for* mixed *strategies easily extend to* asymmetric *preferences (e.g., under semi-bandit feedback, this only doubles the number of utility estimates per agent pair). Yet, extending our analysis for* pure *strategies to* asymmetric *preferences is generally neither theoretically justified nor tractable. Namely, symmetric POCF games always admit a* pure *NS outcome (Lemma 1), but* asymmetric *ones may not, and deciding existence is strongly NP-complete [40].*

4 SEMI-BANDIT FEEDBACK

For semi-bandit feedback settings, we herein derive a ***necessary*** and ***sufficient*** dataset coverage assumption (Section 4.1), under which we devise an offline learning algorithm with low duality gap and a sample complexity bound with ***optimality*** guarantees up to logarithmic factors (Section 4.2).

4.1 The Coalition Size Coverage Assumption

Under semi-bandit feedback, we obtain each agent's individual utility from interacting with every member of her chosen coalitions. This motivates us to require that the dataset covers the kinds of partitions that could arise from agents unilaterally deviating. Specifically, for any $\ell \in [k]$, if the ℓ-th candidate coalition formed by some *unilateral deviation* from a single NS strategy $\boldsymbol{\varphi}^{\star}$ has size $m \in [k]$, then the dataset should contain at least one joint action in which the ℓ-th candidate coalition also has size m. The key intuition is as follows. For any joint action $\mathbf{a} \in \mathcal{A}$ sampled from some joint strategy $\boldsymbol{\varphi}$ and coalition $C \in \pi^{\mathbf{a}}$, consider a unilateral deviation of some agent i. If $i \in C$, then agent i may either remain in or leave C after she deviates. Otherwise, if $i \notin C$, then agent i may either join C or stay out after deviating. Thus, each unilateral deviation affects coalition sizes by either increasing or decreasing them by 1, or leaving them unchanged. This allows us to reason about candidate coalitions and their sizes by considering only such deviations.

Formally, for any joint strategy $\boldsymbol{\varphi}$, $\ell \in [k]$ and coalition size $\alpha \in [n] \cup \{0\}$, let $d_{\ell}^{\boldsymbol{\varphi}}(\alpha) = \sum_{\mathbf{a} \in \mathcal{A}: |C_{\ell}^{\mathbf{a}}| = \alpha} \boldsymbol{\varphi}(\mathbf{a})$ be ***coalitional overall density*** of the ℓ-th candidate coalition, where $d_{\ell}^{\rho}(\alpha)$ is defined similarly for the exploration policy ρ. Our assumption is then:

Assumption 1 (**Coalition Size Coverage**). *There exists an NS joint strategy[4] $\boldsymbol{\varphi}^{\star}$ such that, for any agent i, any $\ell \in [k]$ and any coalition size $\alpha \in [n] \cup \{0\}$, if there is a strategy $\phi_i \in \Delta(\mathcal{A}_i)$ with $d_{\ell}^{\boldsymbol{\varphi}^{\star}_{-i}, \phi_i}(\alpha) > 0$, then $d_{\ell}^{\rho}(\alpha) > 0$.*

Remark 3. *There are realistic scenarios where Assumption 1 can be easily verified. For instance, it is* always *satisfied by an exploration policy that samples joint actions uniformly at random, as such a policy covers all coalitions sizes. Datasets generated by this policy reflect early exploratory behavior, where agents lack prior knowledge.*

We thus measure how well the dataset covers all coalition sizes through unilateral deviations from a joint strategy $\boldsymbol{\varphi}$ via the ***coalition size coefficient***, which is defined by:

$$c_{\mathrm{size}}^{\boldsymbol{\varphi}} = \max_{i \in \mathcal{N}, \ell \in [k], \phi_i \in \Delta(\mathcal{A}_i), \alpha \in [n] \cup \{0\}: d_{\ell}^{\rho}(\alpha) > 0} \frac{d_{\ell}^{\boldsymbol{\varphi}_{-i}, \phi_i}(\alpha)}{d_{\ell}^{\rho}(\alpha)} \tag{7}$$

Next, we prove that Assumption 1 is ***necessary***, i.e., no *weaker* assumption enables efficient learning of an approximate NS mixed strategy with a small duality gap, regardless of the dataset size.

Theorem 2. *Let $\mathcal{G}$ be the class of all pairs (G, ρ) consisting of a POCF game G and an exploration policy ρ satisfying Assumption 1, except for at most one coalition size $\alpha \in [n] \cup \{0\}$. Then, for any algorithm* ALG *with* semi-bandit *feedback, there is $(G, \rho) \in \mathcal{G}$ such that any joint strategy $\boldsymbol{\varphi}$ produced by* ALG *satisfies* $\mathrm{Gap}(\boldsymbol{\varphi}) \geq \frac{1}{2}$ *for the POCF game G, regardless of the dataset size.*

Proof. (*Sketch*) In [12, Appendix D], we construct two symmetric POCF games G_1 and G_2 with 6 agents whose utility distributions are deterministic, where the number of coalitions is at most 2 (i.e., $k = 2$), the action space of each agent is $\{\{1\}, \{2\}\}$ and both games share the same exploration policy ρ. In the first game G_1, there are two types of pure NS joint strategies, which consist of all strategies where either only 2 agents join the first candidate coalition or the grand coalition is formed within that coalition (i.e., all 6 agents join the first candidate coalition). In the second game G_2, there is only one type of pure NS strategies, comprising all strategies where only 5 agents join the first candidate coalition. For both games, we construct the same exploration policy ρ, which picks a joint action uniformly at random from the set of all joint actions where the first

[4] As we focus on *symmetric* POCF games, they always admit a pure NS strategy by Lemma 1, and thus the NS strategy $\boldsymbol{\varphi}^{\star}$ in Assumption 1 can be always chosen as *pure*.

candidate coalition consists of exactly 2, 4 or 5 agents, whereas all other joint actions are assigned zero probability under ρ. Then, we prove that the pairs (G_1, ρ) and (G_2, ρ) both belong to the class $\mathcal{G}$ stated in Theorem 2. Afterwards, we show that any algorithm ALG *cannot* distinguish between $(G_1, \rho), (G_2, \rho)$ as both games appear behaviorally the same from the perspective of the data available under ρ, regardless of the size of the dataset obtained by ALG. Letting q be the probability that a joint action $\mathbf{a}$ inducing a coalition of size 5 is sampled from the joint strategy $\boldsymbol{\varphi}$ produced by ALG, we prove that $\text{Gap}(\boldsymbol{\varphi}) \geq q$ for game G_1 and $\text{Gap}(\boldsymbol{\varphi}) \geq 1 - q$ for game G_2. As either $q \geq \frac{1}{2}$ or $1 - q \geq \frac{1}{2}$, the desired follows. □

4.2 Algorithm 1 under Semi-Bandit Feedback

Next, we prove that Assumption 1 is ***sufficient*** for learning approximate NS strategies. Specifically, we derive the utility estimators and their exploration bonuses for which Algorithm 1 has low duality gap and sample complexity under semi-bandit feedback. Given an offline dataset $\mathcal{S} = \{(\mathbf{a}^m, \mathbf{v}^m)\}_{m=1}^M$, for any $m \in [M]$ and each agent i, recall that $\mathbf{v}^m$ contains agent i's realized utility $v_{i,j}^{\ell,m}$ for each agent $i \neq j \in C_\ell^{\mathbf{a}^m}$, where $C_\ell^{\mathbf{a}^m}$ is the ℓ-th candidate coalition formed by $\mathbf{a}^m$ for some $\ell \in a_i^m$. Hence, we use those feedbacks to estimate each agent's mean utility from each other agent $i \neq j \in \mathcal{N}$ by the empirical average of the utilities she received (if any) from agent j across the entire dataset. Namely, let $N_{i,j}^\ell = \sum_{m=1}^M \mathbb{1}\{\ell \in a_i^m \cap a_j^m\}$ be the number of samples any pair of agents i, j joined the ℓ-th candidate coalition, where $\mathbb{1}\{\ell \in a_i^m \cap a_j^m\}$ equals 1 if $\ell \in a_i^m \cap a_j^m$ and 0 otherwise. Thus, agent i's empirical mean utility from another agent $i \neq j \in \mathcal{N}$ within the ℓ-th candidate coalition is $\hat{v}_i^\ell(j) = \frac{\sum_{m=1}^M v_{i,j}^{\ell,m} \mathbb{1}\{\ell \in a_i^m \cap a_j^m\}}{N_{i,j}^\ell \vee 1}$. Afterwards, we estimate agent i's utility from the partition $\pi^{\mathbf{a}} = (C_\ell^{\mathbf{a}})_{\ell \in [k]}$ induced by any possible joint action $\mathbf{a}$ via:

$$\hat{v}_i(\mathbf{a}) = \sum_{\ell \in a_i} \sum_{i \neq j \in C_\ell^{\mathbf{a}}} \hat{v}_i^\ell(j) \tag{8}$$

To explore a joint action $\mathbf{a}$ more often if it is either promising or not explored enough, we construct agent i's exploration bonus such that it decreases with the increase in $N_{i,j}^\ell$ (here, $\delta \in (0, 1]$ is a confidence level capturing the degree of certainty):

$$b_i^\delta(\mathbf{a}) = \sum_{\ell \in a_i} \sum_{i \neq j \in C_\ell^{\mathbf{a}}} \sqrt{\frac{2\log(4(n+1)k/\delta)}{N_{i,j}^\ell \vee 1}} \tag{9}$$

*4.2.1 Learning Approximate NS **Mixed** Strategies.* As we prove in Theorem 3, we carefully designed (8) and (9) based on Hoeffding's inequality to ensure that Algorithm 1 with the above utility estimators and exploration bonuses has a low duality gap, establishing our algorithm's approximation to Nash stability under *mixed* strategies.

Theorem 3. *Under* semi-bandit *feedback and Assumption 1, for any $\delta \in (0, 1]$, any dataset size $M \in \mathbb{N}$ and any NS joint strategy $\boldsymbol{\varphi}^\star$, the joint mixed strategy $\boldsymbol{\varphi}^{\text{out}}$ formed by Algorithm 1 with utility estimators and exploration bonuses as in* (8)-(9) *satisfies* $\text{Gap}(\varphi^{\text{out}}) \leq \frac{f^\delta(n,k,\boldsymbol{\varphi}^\star)}{\sqrt{M}} + \epsilon_{\text{opt}}$ *with probability at least $1 - \delta$, where:*

$$f^\delta(n, k, \boldsymbol{\varphi}^\star) = 8kn(n+1)c_{\text{size}}^{\boldsymbol{\varphi}^\star} \log\left(\frac{4(n+1)k}{\delta}\right) \sqrt{2(n-1)} \left[\frac{n-1}{2} + \sqrt{\frac{n}{2}}\right] \tag{10}$$

and $c_{\text{size}}^{\boldsymbol{\varphi}^\star}$ is the coalition size coefficient in (7).

Proof. (*Sketch*) In [12, Appendix E], we first use Hoeffding's inequality to prove that b_i as in (9) is an exploration bonus for the utility estimator $\hat{v}_i$ in (8) for any agent i. After bounding $\mathbb{E}_{\mathbf{a} \sim (\boldsymbol{\varphi}_{-i}^\star, \varphi_i')}[b_i^\delta(\mathbf{a})]$ and $\mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}^\star}[b_i^\delta(\mathbf{a})]$ for any strategy $\varphi_i \in \Delta(\mathcal{A}_i)$, we obtain the desired result by Theorem 1. □

For certain values of $\varepsilon > 0$, we now conclude that our algorithm's sample complexity bound for finding an ε-NS strategy ***optimally*** depends on ε up to logarithmic factors.

Corollary 1. *Under* semi-bandit *feedback and Assumption 1, for any $\delta \in (0, 1]$, any $\varepsilon > \epsilon_{\text{opt}}$ with $\epsilon_{\text{opt}} = o(\varepsilon)$ and any NS joint* mixed *strategy $\boldsymbol{\varphi}^\star$, Algorithm 1 with utility estimators and exploration bonuses as in* (8)-(9) *has a sample complexity bound that* **optimally** *depends on ε (up to logarithmic factors): for a dataset of size $M \geq \frac{f^\delta(n,k,\boldsymbol{\varphi}^\star)^2}{(\varepsilon - \epsilon_{\text{opt}})^2}$ (see* (10)*), φ^{out} is ε-NS (i.e., $\text{Gap}(\varphi^{\text{out}}) \leq \varepsilon$) with probability at least $1 - \delta$.*

Proof. (*Sketch*) In [12, Appendix E.2], the desired follows from Theorem 3; our ***optimality*** guarantees w.r.t. ε is by [3, 29]. □

Remark 4. *The scaling with n and k in Theorem 3 and Corollary 1 is* inevitable*: under Assumption 1, to cover all unilateral deviations of n agents across k candidate coalitions, the dataset must cover $\Theta(nk)$ distinct joint actions. Further, the dependence on $c_{\text{size}}^{\boldsymbol{\varphi}^\star}$ is sensible, as its minimum value is at most* 3 *(see [12, Appendix E.1]). Thus, an exploration policy satisfying the* necessary *Assumption 1 with a small $c_{\text{size}}^{\boldsymbol{\varphi}^\star}$ always exists, but it may be hard to find. Empirically, our algorithm still approaches a low approximation to NS for datasets generated by a uniformly random exploration policy (see Section 6).*

*4.2.2 Learning Approximate NS **Pure** Strategies.* As we focus on *symmetric* POCF games, they always admit a pure NS strategy by Lemma 1, which justifies learning an approximate NS *pure* strategy, unlike *asymmetric* preferences (see Remark 2). Hence, we can adapt Algorithm 1 to produce a pure strategy. Indeed, let $\Gamma_i \subset \Delta(\mathcal{A}_i)$ be the set of agent i's *deterministic* strategies, i.e., each deterministic strategy $\varphi_i \in \Gamma_i$ corresponds to exactly one *pure* strategy $a_i \in \mathcal{A}_i$, such that $\varphi_i(a_i) = 1$ for any agent i and $\varphi_i(a_i') = 0$ for any other pure strategy $a_i \neq a_i' \in \mathcal{A}_i$. Thus, instead of approximately solving (6) over *mixed* strategies, Algorithm 1 can be modified to find a joint *deterministic* strategy $\boldsymbol{\varphi}^{\text{out}} \in \Gamma := \prod_{i=1}^n \Gamma_i$ that approximately solves the minimization problem $\min_{\boldsymbol{\varphi} \in \Gamma} \max_{i \in \mathcal{N}} [\overline{V}_i^{\boldsymbol{\star}, \delta}(\boldsymbol{\varphi}_{-i}) - \underline{V}_i^\delta(\boldsymbol{\varphi})]$. See Algorithm 2 in [12, Appendix E.3] for a pseudo-code of the resulting algorithm. The approximation to NS and the sample complexity of the modified Algorithm 1 for *pure* strategies follow from arguments similar to Theorem 3 and Corollary 2 (see Appendices E.3-E.4).

5 BANDIT FEEDBACK

Under bandit feedback, we prove that Assumption 1 is insufficient. Intuitively, as single-agent utilities are unobservable in this setting, we *cannot* estimate an agent's mean utility from any other agent. This may prevent us from accurately estimating utilities obtained from unilateral deviations, as required by Assumption 1. Formally:

Theorem 4. *Let $\mathcal{G}'$ be the class of all pairs (G, ρ) consisting of a POCF game G and an exploration policy ρ satisfying Assumption 1. Then, for any algorithm* ALG *with* bandit *feedback, there is $(G, \rho) \in$*

$\mathcal{G}$ such that any joint strategy $\boldsymbol{\varphi}$ produced by ALG *satisfies* $\mathrm{Gap}(\boldsymbol{\varphi}) \geq \frac{1}{20}$ *for the POCF game G, regardless of the dataset size.*

Proof. (*Sketch*) In [12, Appendix F], we build two symmetric POCF games G_1 and G_2 with 3 agents whose utility distributions are deterministic, where the number of coalitions is at most 3 (i.e., $k = 3$), the action space of each agent is $\{\{1\}, \{2\}, \{3\}, \{1, 2\}\}$, and both games share the same exploration policy ρ. Similarly to Theorem 2, both games are built so that their pure NS joint strategies differ, while ρ is designed such that (G_1, ρ) and (G_2, ρ) are both in the class $\mathcal{G}'$ defined in Theorem 4. We then show that any algorithm ALG *cannot* distinguish between (G_1, ρ), (G_2, ρ), as both games are behaviorally indistinguishable under the data distribution ρ, regardless of the dataset size observed by ALG. By arguments similar to Theorem 2, the desired follows. □

5.1 Algorithm 1 under Bandit Feedback

To circumvent the impossibility in Theorem 4, we next show how to use ridge regression for constructing utility estimators with corresponding exploration bonuses under bandit feedback.[5] To this end, we first show that the mean utility of each agent i from some joint action $\mathbf{a}$ can be written as an inner product between a vector capturing all single-agent utilities and a binary vector, indicating whether agent i and any other agent j join the same coalitions under $\mathbf{a}$. This allows us to derive an assumption on the information covered by the dataset, which is sufficient for sample-efficient learning of an approximate NS strategy with a sample complexity bound that admits ***optimality*** guarantees up to logarithmic factors.

Formally, consider any agent i and any $\ell \in [k]$. We represent agent i's mean mutual utility from interacting with agent j in the ℓ-th candidate coalition via an n-dimensional vector $\boldsymbol{\theta}_i^\ell$, whose j-th coordinate is $[\boldsymbol{\theta}_i^\ell]_j = d_{i,j}^\ell$. We also define a vector-valued function $\mathbf{y}_i^\ell : \mathcal{A} \to \{0,1\}^n$, where $\mathbf{y}_i^\ell(\mathbf{a})$ is an n-dimensional binary vector whose j-th coordinate is 1 if both agents i and j join the ℓ-th candidate coalition under some joint action $\mathbf{a} \in \mathcal{A}$, i.e., $[\mathbf{y}_i^\ell(\mathbf{a})]_j = \mathbb{1}\{\ell \in a_i \cap a_j\}$. We then concatenate agent i's single-agent utilities across coalitions into an nk-dimensional vector $\boldsymbol{\theta}_i = [\boldsymbol{\theta}_i^\ell]_{\ell \in [k]}$, and similarly define $\mathbf{y}_i : \mathcal{A} \to \{0,1\}^{nk}$ via $\mathbf{y}_i(\mathbf{a}) = [\mathbf{y}_i^\ell(\mathbf{a})]_{\ell \in [k]}$. Further, we concatenate all agents' utilities into an n^2k-dimensional vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_i]_{i \in \mathcal{N}}$. Finally, letting $\mathbf{0}_\kappa$ be the κ-dimensional all-zeros vector, we denote the concatenation $\mathbf{z}_i(\mathbf{a}) = [\mathbf{0}_{k(i-1)}, \mathbf{y}_i(\mathbf{a}), \mathbf{0}_{k(n-i+1)}]$.

As $d_i(\mathbf{a}) = \sum_{\ell \in a_i} \sum_{i \neq j \in C_\ell^{\mathbf{a}}} d_{i,j}^\ell$ is agent i's mean utility from a joint action $\mathbf{a} \in \mathcal{A}$, then it can be written as $d_i(\mathbf{a}) = \langle \mathbf{z}_i(\mathbf{a}), \boldsymbol{\theta} \rangle$. Therefore, under bandit feedback, we can construct a utility estimator $\hat{v}_i$ for agent i's mean utility with an exploration bonus b_i^δ for a confidence level $\delta \in (0, 1]$ using *ridge regression* over a dataset $\mathcal{S} = \{(\mathbf{a}^m, \mathbf{v}^m)\}_{m=1}^M$ to estimate $\boldsymbol{\theta}$ as follows:

$$\begin{aligned} \hat{v}_i(\mathbf{a}) &= \langle \mathbf{z}_i(\mathbf{a}), \hat{\boldsymbol{\theta}} \rangle \\ b_i^\delta(\mathbf{a}) &= \|\mathbf{z}_i(\mathbf{a})\|_{V^{-1}} \sqrt{\beta} = \sqrt{\mathbf{z}_i(\mathbf{a})^\top V^{-1} \mathbf{z}_i(\mathbf{a}) \beta} \\ \hat{\boldsymbol{\theta}} &= V^{-1} \sum_{m \in [M]} \sum_{i \in \mathcal{N}} \mathbf{z}_i(\mathbf{a}^m) v_i(\mathbf{a}^m) \\ V &= I + \sum_{m \in [M]} \sum_{i \in \mathcal{N}} \mathbf{z}_i(\mathbf{a}^m) \mathbf{z}_i(\mathbf{a}^m)^\top \\ \sqrt{\beta} &= 2\sqrt{n^2 k} + \sqrt{n^2 k \log(1 + M/n) + \iota} \end{aligned} \tag{11}$$

[5] Our results in this section extend to learning approximate NS **pure** strategies by arguments similar to Section 4.2.2, and are thus deferred to [12, Appendix G.3].

where $\iota = 2\log(4(n+1)k/\delta)$. Next, we derive our assumption on the information contained in the dataset. Intuitively, for any agent i and any unilateral deviation of agent i from some NS joint strategy, it ensures that the dataset contains at least as much information as a sufficiently large dataset generated according to the joint strategy induced by that unilateral deviation, so that the associated utilities can be reliably estimated. Formally:

Assumption 2 (**Action Coverage**). *There exist a universal constant $c_{\mathrm{act}} > 0$ and an NS joint strategy $\boldsymbol{\varphi}^\star$ such that, for any agent i and any strategy $\phi_i \in \Delta(\mathcal{A}_i)$, it holds that:*

$$V \succeq I + M c_{\mathrm{act}} \mathbb{E}_{\mathbf{a} \sim (\boldsymbol{\varphi}_{-i}^\star, \phi_i)} [\mathbf{z}_i(\mathbf{a}) \mathbf{z}_i(\mathbf{a})^\top] \tag{12}$$

Here, V is the covariance matrix of the dataset as defined in (11), while the right-hand side of (12) is the covariance matrix of joint actions sampled over Mc_{act} episodes from the strategy $(\boldsymbol{\varphi}_{-i}^\star, \phi_i)$ induced by a unilateral deviation $\phi_i \in \Delta(\mathcal{A}_i)$ from an NS joint strategy $\boldsymbol{\varphi}^\star$. Thus, to enable accurate utility estimation, Assumption 2 requires that the dataset is at least as informative as a dataset of Mc_{act} samples independently drawn from $(\boldsymbol{\varphi}_{-i}^\star, \phi_i)$.

Under Assumption 2, we are now ready to prove that Algorithm 1 with utility estimators and exploration bonuses as in (11) has a low duality gap. In particular, Theorem 5 quantifies the approximation guarantees to Nash stability obtained by Algorithm 1 with utility estimators and exploration bonuses as in (11). Formally:

Theorem 5. *Under* bandit *feedback and Assumption 2, for any $\delta \in (0, 1]$ and dataset size $M \in \mathbb{N}$, Algorithm 1 with utility estimators and exploration bonuses as in* (11) *outputs a joint* mixed *strategy $\boldsymbol{\varphi}^{\mathrm{out}}$ satisfying* $\mathrm{Gap}(\varphi^{\mathrm{out}}) \leq 4\sqrt{\frac{n^2 k \beta}{c_{\mathrm{act}} M}} + \epsilon_{\mathrm{opt}}$ *with probability at least $1 - \delta$.*

Proof. (*Sketch*) In [12, Appendix G], we prove that b_i^δ in (11) is an exploration bonus for the utility estimator $\hat{v}_i$ in (11) for any agent i. After bounding $\mathbb{E}_{\mathbf{a} \sim (\boldsymbol{\varphi}_{-i}^\star, \varphi_i')}[b_i^\delta(\mathbf{a})]$ and $\mathbb{E}_{\mathbf{a} \sim \boldsymbol{\varphi}^\star}[b_i^\delta(\mathbf{a})]$ by $\sqrt{\frac{n^2 k \beta}{c_{\mathrm{act}} M}}$ for any $\varphi_i \in \Delta(\mathcal{A}_i)$, we obtain the desired result by Theorem 1. □

For specific values of $\varepsilon > 0$, we now infer that our algorithm's sample complexity bound for reaching an ε-NS strategy ***optimally*** depends on ε up to logarithmic factors. The proof is by arguments similar to Corollary 1, and thus deferred to [12, Appendix G.2].

Corollary 2. *Under* bandit *feedback and Assumption 2, for any $\delta \in (0, 1]$, any $\varepsilon > \epsilon_{\mathrm{opt}}$ with $\epsilon_{\mathrm{opt}} = o(\varepsilon)$ and any NS joint strategy $\boldsymbol{\varphi}^\star$, Algorithm 1 with utility estimators and exploration bonuses as in* (11) *has a sample complexity bound that* **optimally** *depends on ε (up to logarithmic factors): for a dataset of size $M \geq \frac{16 n^2 k \beta}{c_{\mathrm{act}} (\varepsilon - \epsilon_{\mathrm{opt}})^2}$, φ^{out} is ε-NS (i.e., $\mathrm{Gap}(\varphi^{\mathrm{out}}) \leq \varepsilon$) with probability at least $1 - \delta$.*

Remark 5. *The scaling with n and k in Theorem 5 and Corollary 2 is* unavoidable*: Assumption 2 requires that the dataset's covariance matrix is as informative as if we had sampled from any of the $\Theta(nk)$ possible unilateral deviations. Further, the dependence on c_{act} is reasonable: even in games where each agent i can join any possible non-empty subset of candidate coalitions (i.e., $|\mathcal{A}_i| = 2^k - 1$), we prove in [12, Appendix G.1] that there is always a sufficiently large dataset such that Assumption 2 holds for $c_{\mathrm{act}} = \frac{1}{2nk^4}$ with high probability.*

6 EMPIRICAL EVALUATIONS

We evaluate Algorithm 1 through extensive experiments on several synthetic datasets [13]. Due to space constraints, we herein focus on *semi-bandit* feedback, with *bandit* feedback results deferred to [12, Appendix H.2]. Our experiments evaluate how well our algorithm recovers approximate NS outcomes across various settings.

Setup. For each run, we generate a game with n agents who share the same action set of size at least 3, sampled uniformly at random from $\mathcal{P}([k]) \setminus \emptyset$. Given a joint action $\mathbf{a}$, we sample the mutual utility $v_{i,j}^{\ell}$ of each pair of distinct agents i, j in the ℓ-th candidate coalition $C_{\ell}^{\mathbf{a}}$ using one of the following two utility generation models inspired by Boehmer et al. [7]: (1) **Size-Dependent Uniform:** We first draw $u_{i,j}^{\ell}$ uniformly at random from $[-1, 1]$ and then set $v_{i,j}^{\ell} = \frac{|C_{\ell}^{\mathbf{a}}|}{n+1} \cdot u_{i,j}^{\ell}$; (2) **Size-Dependent Gaussian:** We first draw a mean $\mu_{i,j}$ uniformly at random from $[-1, 1]$. Then, $u_{i,j}^{\ell}$ is drawn from the Gaussian distribution with mean $\mu_{i,j}$ and standard deviation $1 - \mu_{i,j}$ if $\mu_{i,j} \geq 0$ and $|-1 - \mu_{i,j}|$ if $\mu_{i,j} < 0$, ensuring that $v_{i,j}^{\ell} \in [-1, 1]$. Finally, we set $v_{i,j}^{\ell} = \frac{|C_{\ell}^{\mathbf{a}}|}{n+1} \cdot u_{i,j}^{\ell}$. We also considered size-independent and mixed size effects variants, which showed similar trends to the size-dependent versions; their results are thus deferred to [12, Appendix H.1].

For each configuration of parameters, we consider two exploration policies for constructing an offline dataset of size M: (1) The ***uniformly random*** policy ρ^{rand}, where joint actions are sampled uniformly at random (i.e., $\rho^{\text{rand}}(\mathbf{a}) = 1/|\mathcal{A}|$ for any joint action $\mathbf{a} \in \mathcal{A}$). ρ^{rand} covers all coalition sizes, thus satisfying Assumption 1; (2) To validate the need of Assumption 1, we also consider a policy ρ^{1Rand} that does not necessarily satisfy it, which induces a uniformly random strategy ρ_1^{1Rand} for agent 1 (i.e., $\rho_1^{\text{1Rand}}(a_1) = \frac{1}{|\mathcal{A}_1|}$ for any $a_1 \in \mathcal{A}_1$); others always deterministically follow the second action that was inserted to their action set during its random generation.

Algorithm 1 under semi-bandit feedback then uses the exploration bonuses in (9) with confidence level $\delta = 10^{-2}$. As exactly solving (6) in Algorithm 1 is intractable (Footnote 3), we employ a practical *coordinate-descent* scheme, widely used in large-scale optimization for its efficient convergence properties to stationary points (e.g., [33, 45]). At each round, we update each agent's mixed strategy via a convex combination of her current strategy and an optimistic best response to the estimated empirical utilities of the other agents. We stop once improvements fall below 10^{-3}. This yields a smoothed best-response dynamics with gradual convergence to per-agent optimistic best responses. Here, all expectations are estimated by Monte Carlo sampling with 100 samples per term.

Results. Fig. 1 reports the mean approximate duality gap $\widehat{\text{Gap}}^{\delta}(\boldsymbol{\varphi}^{\text{out}})$ of the strategy $\boldsymbol{\varphi}^{\text{out}}$ produced by Algorithm 1 versus the size of datasets generated by ρ^{rand} (top two rows) and ρ^{1Rand} (last row) over 5 runs with different seeds. By Lemma 2, $\widehat{\text{Gap}}^{\delta}(\boldsymbol{\varphi}^{\text{out}})$ upper bounds the *true* duality gap with high probability, thus quantifying how close the learned strategy is to Nash stability. The left column examines the effect of varying the number of agents $n \in \{5, 10, 15, 20, 25\}$ while fixing $k = 5$, whereas the right column varies the number of candidate coalitions $k \in \{5, 10, 15, 20, 25\}$ while fixing $n = 10$. In each experiment, we evaluate our algorithm on datasets of sizes $M \in \{10^2, 5 \cdot 10^3, 10^4, 2 \cdot 10^4, 3 \cdot 10^4\}$. For ρ^{1Rand}, we report only size-independent uniform utilities; the Gaussian variant shows similar trends and is thus deferred to [12, Appendix H.1].

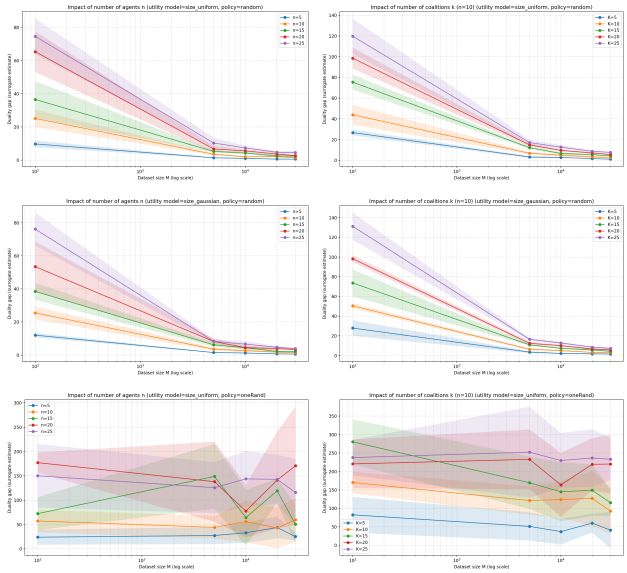

Figure 1: Mean approximate duality gap versus the size of datasets generated by ρ^{rand} (top two rows) and ρ^{1Rand} (last row) over 5 runs with different seeds, for varying numbers of agents (left column) and candidate coalitions (right column). Shaded regions indicate standard deviations.

Algorithm 1 consistently reaches a low approximation to Nash stability under ρ^{rand}, but fails to do so under ρ^{1Rand}. This supports the practical relevance of Assumption 1: ρ^{rand} satisfies it, allowing effective learning, whereas ρ^{1Rand} may violate it, yielding poorer performance. For ρ^{rand}, the scaling in n, k, M also matches Theorem 3, Corollary 1 and Remark 4. For any fixed number of agents n, the gap decreases rapidly when the dataset size M is small, but slowly for larger M, consistent with the $1/\sqrt{M}$ factor in Theorem 3. Further, obtaining a certain approximation requires larger datasets as n increases, in line with Corollary 1 and Remark 4.

7 CONCLUSIONS AND FUTURE WORK

We presented a new model for studying coalition formation with *possibly overlapping* coalitions under *partial information*, where agents' preferences must be inferred from a fixed offline dataset. Under both semi-bandit and bandit feedback, we identified conditions under which the dataset covers sufficient information for an offline learning algorithm to infer preferences and use them to recover an approximately NS joint strategy. Under those conditions, we designed sample-efficient algorithms whose sample complexity bounds for learning an ε-approximate NS strategy ***optimally*** depend on ε up to logarithmic factors for certain values of $\varepsilon > 0$.

Our research offers many promising directions for future works. Immediate directions are exploring additional types of preferences, solution concepts and models of partial and/or noisy information. Finally, while we consider datasets with independent samples, common in offline learning (see, e.g., [20, 21]), studying *correlated* samples remains an open challenge, even in single-agent offline reinforcement learning (see, e.g., [20, 37]).

Acknowledgments. The author acknowledges travel support from a Schmidt Sciences 2025 Senior Fellows award to Michael Wooldridge.

REFERENCES

[1] Haris Aziz, Felix Brandt, and Hans Georg Seedig. 2011. Stable partitions in additively separable hedonic games.. In *AAMAS*, Vol. 11. 183–190.

[2] Haris Aziz and Rahul Savani. 2016. Hedonic Games. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D.Editors Procaccia (Eds.). Cambridge University Press, 356–376.

[3] Yu Bai and Chi Jin. 2020. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *International Conference on Machine Learning*, Vol. 119. PMLR, 551–560.

[4] Maria-Florina Balcan, Ariel D Procaccia, and Yair Zick. 2015. Learning cooperative games. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 475–481.

[5] Coralio Ballester. 2004. NP-completeness in hedonic games. *Games and Economic Behavior* 49, 1 (2004), 1–30.

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

[7] Niclas Boehmer, Martin Bullinger, and Anna Maria Kerkmann. 2025. Causes of stability in dynamic coalition formation. *ACM Transactions on Economics and Computation* 13, 2 (2025), 1–45.

[8] Anna Bogomolnaia and Matthew O Jackson. 2002. The stability of hedonic coalition structures. *Games and Economic Behavior* 38, 2 (2002), 201–230.

[9] Martin Bullinger and René Romen. 2025. Stability in online coalition formation. *Journal of Artificial Intelligence Research* 82 (2025), 2423–2452.

[10] Georgios Chalkiadakis, Edith Elkind, Evangelos Markakis, Maria Polukarov, and Nick R Jennings. 2010. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research* 39 (2010), 179–216.

[11] Saar Cohen. 2026. Delayed Assignments in Online Non-Centroid Clustering with Stochastic Arrivals. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*.

[12] Saar Cohen. 2026. Offline Learning of Nash Stable Coalition Structures with Possibly Overlapping Coalitions. *arXiv preprint arXiv:2602.14321* (2026).

[13] Saar Cohen. 2026. Offline Learning of Nash Stable Coalition Structures with Possibly Overlapping Coalitions. https://github.com/saarcohen30/pocf. In AAMAS'26: The 25th International Conference on Autonomous Agents and Multiagent Systems.

[14] Saar Cohen and Noa Agmon. 2023. Online Coalitional Skill Formation. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS*. 494–503.

[15] Saar Cohen and Noa Agmon. 2024. Online Friends Partitioning Under Uncertainty. In *ECAI 2024*. IOS Press, 3332–3339.

[16] Saar Cohen and Noa Agmon. 2024. Online Learning of Partitions in Additively Separable Hedonic Games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. 2722–2730.

[17] Saar Cohen and Noa Agmon. 2025. Decentralized online learning by selfish agents in coalition formation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 3796–3804.

[18] Saar Cohen and Noa Agmon. 2025. Egalitarianism in online coalition formation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2475–2477.

[19] Saar Cohen and Noa Agmon. 2025. Online Learning of Coalition Structures by Selfish Agents. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 13 (2025), 13709–13717.

[20] Qiwen Cui and Simon S Du. 2022. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Advances in Neural Information Processing Systems* 35 (2022), 11739–11751.

[21] Qiwen Cui and Simon S Du. 2022. When are offline two-player zero-sum Markov games solvable? *Advances in Neural Information Processing Systems* 35 (2022), 25779–25791.

[22] Viet Dung Dang, Rajdeep K Dash, Alex Rogers, and Nicholas R Jennings. 2006. Overlapping coalition formation for efficient data fusion in multi-sensor networks. In *AAAI*, Vol. 6. 635–640.

[23] Dinko Dimitrov, Peter Borm, Ruud Hendrickx, and Shao Chin Sung. 2006. Simple priorities and core stability in hedonic games. *Social Choice and Welfare* 26, 2 (2006), 421–433.

[24] Jacques H Dréze and Joseph Greenberg. 1980. Hedonic coalitions: Optimality and stability. *Econometrica: Journal of the Econometric Society* (1980), 987–1003.

[25] Edith Elkind, Talal Rahwan, and Nicholas R Jennings. 2013. Computational coalition formation. *Multiagent systems* (2013), 329–380.

[26] Simone Fioravanti, Michele Flammini, Bojana Kodric, and Giovanna Varricchio. 2023. PAC learning and stabilizing Hedonic Games: towards a unifying approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (2023), 5641–5648.

[27] Michele Flammini, Bojana Kodric, Gianpiero Monaco, and Qiang Zhang. 2021. Strategyproof mechanisms for additively separable and fractional hedonic games. *Journal of Artificial Intelligence Research* 70 (2021), 1253–1279.

[28] Martin Gairing and Rahul Savani. 2019. Computing stable outcomes in symmetric additively separable hedonic games. *Mathematics of Operations Research* 44, 3 (2019), 1101–1121.

[29] Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Zebang Shen. 2020. Stochastic conditional gradient++: (non)convex minimization and continuous submodular maximization. *SIAM Journal on Optimization* 30, 4 (2020), 3315–3344.

[30] Ayumi Igarashi, Jakub Sliwinski, and Yair Zick. 2019. Forming probably stable communities with limited interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2053–2060.

[31] Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021. Is pessimism provably efficient for offline RL?. In *International conference on machine learning*. PMLR, 5084–5096.

[32] Chao-Feng Lin and Shan-Li Hu. 2007. Multi-task overlapping coalition parallel formation algorithm. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 1–3.

[33] Nesterov, Yu. 2012. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization* 22, 2 (2012), 341–362.

[34] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. 2021. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems* 34 (2021), 11702–11716.

[35] Onn Shehory and Sarit Kraus. 1996. Formation of overlapping coalitions for precedence-ordered task-execution among autonomous agents. In *Proceedings of international conference on multi agent systems*. CiteSeer, 330–337.

[36] Onn Shehory and Sarit Kraus. 1998. Methods for task allocation via agent coalition formation. *Artificial intelligence* 101, 1-2 (1998), 165–200.

[37] Shi, Chengshuai and Xiong, Wei and Shen, Cong and Yang, Jing. 2023. Provably efficient offline reinforcement learning with perturbed data sources. In *International Conference on Machine Learning*. PMLR, 31353–31388.

[38] Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. 2020. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2992–3002.

[39] Jakub Sliwinski and Yair Zick. 2017. Learning Hedonic Games. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17*. 2730–2736.

[40] Shao-Chin Sung and Dinko Dimitrov. 2010. Computational complexity in additive hedonic games. *European Journal of Operational Research* 203, 3 (2010), 635–639.

[41] Csaba Szepesvári. 2022. *Algorithms for reinforcement learning*. Springer nature.

[42] Csaba Szepesvári and Rémi Munos. 2005. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*. 880–887.

[43] Ruosong Wang, Dean Foster, and Sham M Kakade. 2021. What are the Statistical Limits of Offline RL with Linear Function Approximation?. In *International Conference on Learning Representations*.

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

[45] Wright, Stephen J. 2015. Coordinate descent algorithms. *Mathematical programming* 151, 1 (2015), 3–34.

[46] Ming Yin and Yu-Xiang Wang. 2021. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems* 34 (2021), 4065–4078.

[47] Guofu Zhang, Jianguo Jiang, Zhaopin Su, Meibin Qi, and Hua Fang. 2010. Searching for overlapping coalitions in multiple virtual organizations. *Information Sciences* 180, 17 (2010), 3140–3156.

[48] Guofu Zhang, Zhaopin Su, Miqing Li, Meibin Qi, Jianguo Jiang, and Xin Yao. 2017. A task-oriented heuristic for repairing infeasible solutions to overlapping coalition structure generation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 50, 3 (2017), 785–801.

[49] Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. 2023. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Journal of Machine Learning Research* 24, 175 (2023), 1–53.

[50] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. 2021. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Trans. Automat. Control* 66, 12 (2021), 5925–5940.

[51] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. 2022. Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. 27117–27142.

[52] Yair Zick, Georgios Chalkiadakis, Edith Elkind, and Evangelos Markakis. 2019. Cooperative games with overlapping coalitions: Charting the tractability frontier. *Artificial Intelligence* 271 (2019), 74–97.

[53] Yair Zick, Evangelos Markakis, and Edith Elkind. 2012. Stability via convexity and LP duality in OCF games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1506–1512.