

Stability-Accuracy Tradeoffs in Decentralized Online Learning of Coalition Structures by Selfish Agents

SAAR COHEN*, Department of Computer Science, Bar-Ilan University, Israel and Department of Computer Science, University of Oxford, United Kingdom

NOA AGMON, Department of Computer Science, Bar-Ilan University, Israel

Coalition formation involves self-organized coalitions generated through strategic collaborations of autonomous selfish agents. In *online learning* of coalition structures, agents' preferences toward each other are initially unknown *before* agents interact. Coalitions are formed iteratively based on preferences that agents learn *online* from repeated feedback resulting from their interactions. In this work, we introduce online learning in coalition formation through the lens of *distributed decision-making*, where self-interested agents operate without global coordination or information sharing, and learn only from their own experience. Under the selfish perspective, each agent seeks to maximize her own utility. Hence, we analyze the system in terms of *Nash stability*, where no agent can improve her utility by unilaterally deviating. Our goal is therefore designing sample-efficient decentralized algorithms for selfish agents that minimize their Nash regret, yielding approximately Nash stable solutions. This poses a natural tension between an algorithm's performance in the short- and long-term, trading off between the accuracy of the utility estimation at each round and the long-term objective of minimizing Nash regret to reach approximately Nash stable outcomes. Specifically, we first devise an algorithm where each agent sacrifices accuracy at each single time step by using only *one* utility feedback per round to update her strategy, but still has Nash regret and sample complexity bounds that are *optimal* up to logarithmic factors. To control the error incurred at any individual time instant, we then propose a method using *multiple* samples per round, yet at the expense of obtaining *suboptimal* Nash regret and sample complexity bounds.

JAIR Associate Editor: Paolo Turrini

JAIR Reference Format:

Saar Cohen and Noa Agmon. 2026. Stability-Accuracy Tradeoffs in Decentralized Online Learning of Coalition Structures by Selfish Agents. *Journal of Artificial Intelligence Research* 87, Article 10 (October 2026), 46 pages. doi: [10.1613/jair.1.19776](https://doi.org/10.1613/jair.1.19776)

1 Introduction

Freelance developers collaborating on open-source projects pursue individual goals like skill enhancement and portfolio building. Their self-interest affects their collaborative choices of forming project teams, but they cannot make optimal decisions as they initially lack clarity on their preferences about other developers, team size, or project types. After contributing to a certain team, these selfish developers learn about their preferences from feedback solely about their own team's productivity and satisfaction with the collaborative dynamics. This adaptive process allows them to selfishly optimize their working experiences, making informed decisions to join project teams that best align with their personal goals and working styles. Such scenarios and many other

*Corresponding Author.

Authors' Contact Information: Saar Cohen, ORCID: [0000-0001-8262-405X](https://orcid.org/0000-0001-8262-405X), saar30@gmail.com, Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel and Department of Computer Science, University of Oxford, Oxford, United Kingdom; Noa Agmon, ORCID: [0000-0003-0890-7734](https://orcid.org/0000-0003-0890-7734), agmon@cs.biu.ac.il, Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2026 Copyright held by the owner/author(s).

DOI: [10.1613/jair.1.19776](https://doi.org/10.1613/jair.1.19776)

real-life cases exemplify *online learning in coalition formation*, where coalitions are formed iteratively based on preferences that agents learn *online* from repeated feedback about their interactions.

Hedonic games (Dr  ze and Greenberg 1980) are a popular framework for studying coalition formation, where the utilities of *selfish* agents only depend on the coalition they are part of, disregarding the structure of other coalitions, i.e., *externalities* are ignored. The outcome of such games is a set of disjoint coalitions (hereafter, *partition*), whose desirability is often assessed in terms of *stability* (Aziz and Savani 2016; Bullinger and Romen 2025b), reflecting the likelihood of selfish agents maintaining their coalitions. Traditional literature on stability in hedonic games often focuses only on the final outcome of coalition formation, ignoring the process of reaching stable partitions. Particularly, in most existing works it is implicitly assumed that a *central* authority can attain the agents' preferences, find a stable partition and impose it on the agents. Recent works, however, consider a *dynamic* process where, starting from an initial partition, agents deliberately move between coalitions based on their preferences (Boehmer et al. 2023; Brandt, Bullinger, and Tappe 2022). In many realistic cases as our project teams example, agents' preferences toward others are initially *unknown* prior to interactions (see, e.g., (Cohen and Agmon 2023a)). Hence, each agent must make decisions individually based on her own preferences, which she learns *online* from repeated feedback by iteratively joining coalitions. To reflect such cases, Cohen and Agmon (Cohen and Agmon 2024b) study online learning in hedonic games, aiming to maximize social welfare. Recently, Cohen and Agmon (Cohen and Agmon 2025b) explored a *centralized* setting where *selfish* agents seek to maximize their own utility.

In contrast, in this paper we introduce and study a *new* model for online learning in coalition formation, reflecting such realistic situations from the perspective of *decentralized decision-making* by self-interested agents. As agents are usually not aware of other agents' strategies, we assume *bandit feedback*, i.e., agents solely observe their utility from the coalition they joined. We exhibit our findings for *additively separable hedonic games* with *symmetric* preferences (Bogomolnaia and Jackson 2002), where an agent's utility for a coalition is the sum of her utilities from other coalition members. Our framework has no global coordination among agents. Therefore, agents learn only from their own history of strategies, coalitions' composition and utility feedbacks.

From our selfish viewpoint, we evaluate partitions by means of *Nash stability*, where no agent can improve her utility by unilaterally deviating. This stability notion is well-suited to our distributed context as it does not allow any agent to coordinate with others to determine if she can increase her utility (see, e.g., (Balliu et al. 2019)). In this setting, the partition formed by uncoordinated selfish agents can be inferior to a centrally designed one. Under our online learning setup, a popular metric for such inefficiency of equilibria is *Nash regret* (Ding et al. 2022; Q. Liu et al. 2021), comparing an agent's learnt strategy against her best response strategy at any round. Sublinear Nash regret translates to low sample complexity as it implies best-iterate convergence to an approximate Nash stable solution with fewer samples (by classic online-to-batch conversion due to, e.g., (Jin et al. 2018)), strongly tying these notions to our online learning setting.

Our Contributions. We develop sample-efficient decentralized algorithms for online learning in coalition formation by selfish agents that minimizes its Nash regret, thus obtaining approximately Nash stable partitions. Such a distributed approach is also often favored over centralized ones due to its simplicity of implementation, versatility, and faster execution resulting from reduced communication overhead. This setting naturally raises a tradeoff between the short-term goal of optimizing the accuracy of utility estimation at each round and the long-term objective of minimizing Nash regret to obtain approximately Nash stable outcomes. Particularly, we devise two Frank-Wolfe algorithms for *distributed* online learning in coalition formation, which are projection-free as they replace the projections used by prior methods with a more efficient linear optimization step. In our first algorithm, each agent compromises accuracy at each time step by only utilizing *one* sample per round to update her strategy, yet it obtains Nash regret and sample complexity bounds that are *optimal* (up to logarithmic

factors). To reduce the error incurred at each single round, we then develop a method using *multiple* samples per round, but yielding *suboptimal* Nash regret and sample complexity bounds¹.

2 Related Work

Hedonic games were introduced by Dr  ze and Greenberg (1980), and later expanded to the study of various notions of stability, fairness, and optimality (see, e.g., (Aziz and Savani 2016)). We focus on *additively separable hedonic games* (ASHGs) with *symmetric* preferences (Bogomolnaia and Jackson 2002), where a large body of work evaluates the system in terms of *stability*. Unlike *cooperative* approaches (e.g., core stability (Bogomolnaia and Jackson 2002)), we explore a *non-cooperative* perspective, where many works study Nash stability (Aloisio et al. 2020; Ballester 2004; Balliu et al. 2019; Banerjee et al. 2001; Bloch and Diamantoudi 2011; Feldman et al. 2015). Particularly, Bogomolnaia and Jackson (2002) proved that Nash stable partitions may not exist in general ASHGs, while Sung and Dimitrov (2010) showed that checking if an instance admits such a partition is NP-complete in the strong sense. Yet, for *symmetric* preferences, the existence of a Nash stable outcome is guaranteed by potential function argument (Bogomolnaia and Jackson 2002), but computing such partitions is PLS-complete (Gairing and Savani 2019). However, the above works explore *offline* settings, while we regard *online* ones.

Hence, our work is closely related to *time-dependent* models in hedonic games, including their *online* variant introduced by Flammini, Monaco, et al. (2021), where agents arrive one at a time and should be *immediately* and *irrevocably* assigned to coalitions with the goal of maximizing social welfare. This problem was recently extended to other setups (Bullinger and Romen 2025a; Bullinger, Romen, and Schlenga 2026; Cohen 2026a; Cohen and Agmon 2025a, 2024a, 2025b,c, 2024b), with Bullinger and Romen (2025b) also exploring various *stability* concepts. However, the assumption that the agents are partitioned by a *central* authority may be unrealistic, as agents typically make decisions individually. In contrast, we adopt a *decentralized* perspective, where agents selfishly decide to either form a new coalition or to join an existing one based only on their *local* information from past rounds. Recently, dynamic and distributed approaches to hedonic games have also received increased attention, focusing on deviation dynamics (Bil   et al. 2018; Boehmer et al. 2023; Brandt, Bullinger, and Tappe 2022). Related learning-based approaches have also been considered in applied coalition-formation settings; for example, Drechsler (2025) studies coalition formation among learning landowners under an agglomeration-bonus scheme.

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. Unlike our work, this line of research primarily focuses on a *cooperative* approach (Balcan et al. 2015; Cohen 2026b; Fioravanti et al. 2023; Igarashi et al. 2019; Sliwinski and Zick 2017), while Cohen (2026b) analyzes Nash stability. In contrast to our paper, prior studies on PAC learnability in hedonic games aims 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, as in our project teams example. 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. We consider situations where each agent *dynamically* learns her own preferences through repeated interactions, enabling her to adapt to changing scenarios so as to learn the coalitions proven most relevant and effective for her selfish desires.

We propose a novel framework that addresses these challenges by examining online learning in coalition formation through the lens of *decentralized decision-making* by selfish agents, unlike the *centralized* approaches presented by Cohen and Agmon (Cohen and Agmon 2025b, 2024b). We also contribute to the growing focus

¹A preliminary version of this work has been accepted to IJCAI 2025. This submission substantially extends it in two ways: (a) We have significantly improved the presentation and included all proofs that were omitted in the conference version. (b) Unlike our IJCAI paper which focuses solely on a *decentralized one-sample* algorithm, we herein examine the balance between an algorithm's performance in the short- and long-term by introducing an additional *decentralized multi-sample* algorithm.

on online learning in combinatorial domains (e.g., online task allocation (Cohen and Agmon 2023b)). Our work also relates to decentralized multi-agent and multiplayer bandits (Boursier and Perchet 2024; Xu and Klabjan 2025), where agents learn from limited local feedback and regret is a central performance criterion. In contrast, our agents endogenously form coalitions, and we evaluate their joint learning process through Nash regret and approximate Nash stability. Unlike this literature, we study decentralized online learning in coalition formation and evaluate performance through Nash regret. Traditional literature on learning in games often studies how various dynamics asymptotically converge to a Nash equilibrium (e.g., no-regret dynamics (Chen and Peng 2020; Daskalakis, Fishelson, et al. 2021), fictitious play (Daskalakis and Pan 2014; Leslie and Collins 2006)). In contrast, we focus on non-asymptotic convergence, as done in recent studies on multi-agent reinforcement learning where *Nash regret* serves as a key performance metric (Ding et al. 2022; Q. Liu et al. 2021), comparing an agent's learnt strategy with her best response strategy at each round. This metric is thus well-suited to the inherently non-stationary nature of our setting. By classic online-to-batch conversion (e.g., (Jin et al. 2018, Section 3.1)), sublinear Nash regret yields low sample complexity, leading to best-iterate convergence to approximate Nash stable solutions with fewer samples.

Matching markets also relate to our work. Unlike prior works (L. T. Liu et al. 2020; Maheshwari et al. 2022; Y. Zhang et al. 2022), we address more practical and complex settings with broader utility functions under bandit feedback that go beyond matchings. Our work also ties to potential games, where prior distributed methods often rely on computationally expensive projection-based approaches (Ding et al. 2022; Leonardos et al. 2022), using costly or even intractable projection operations. Conversely, our projection-free Frank-Wolfe method resolves both issues by using a more efficient linear optimization step. The sample complexity and Nash regret of prior methods for potential games scale linearly in the number of actions an agent may take, which may be exponential in the number of agents. In contrast, our algorithms remove this dependency while attaining *optimal* sample complexity and Nash regret bounds (up to logarithmic factors). For congestion games, a subclass of potential games, Cui et al. (2022) also provide a projection-free Frank-Wolfe method, yet it suffers from suboptimal sample complexity and Nash regret bounds, relying on large mini-batches of samples and thus not updating optimization parameters frequently enough due to Defazio and Bottou (2019). However, existing one-sample Frank-Wolfe schemes either exhibit suboptimal convergence rates (Mokhtari et al. 2018) or make restrictive assumptions such as knowing each agent's true utility function (M. Zhang et al. 2020). In contrast, without knowing actual utilities, our one-sample algorithm uses a single sample per round while having *optimal* sample complexity and Nash regret bounds (up to logarithmic factors). Our algorithms also contrast with that of Dadi et al. (2024), which has *suboptimal* sample complexity and generally *exponential* running time, becoming polynomial only for restricted cases, whereas our algorithms can be executed in *polynomial* time. Finally, unlike prior methods that update policies *independently* (Ding et al. 2022; Leonardos et al. 2022), our algorithms exploit awareness of other coalition members for better guarantees. For summary, see Table 1 for a comparison of our proposed decentralized algorithms with the aforementioned algorithms after applying them to our setting. Table 1 highlights that, in practice, our one-sample algorithm is expected to theoretically outperform existing decentralized methods originally developed for potential and congestion games.

3 Preliminaries

We study an *online learning* version of hedonic games, where selfish agents with initially *unknown* preferences partition themselves into disjoint subsets (i.e., *coalitions*) over a known number of rounds T , which holds in many real-life scenarios as our project teams example, where developers may join project teams over a predefined number of project milestones. Formally, our strategic game is given by a finite set $N = \{1, \dots, n\}$ of n selfish agents with *unknown* preferences. Hereafter, we denote $[k] := \{1, \dots, k\}$ for $k \in \mathbb{N}$ and $[0] = \{0\}$. At any time $t \in [T]$, each agent can join one of n *candidate* coalitions since there are n agents (i.e., a partition can contain

Table 1. Comparison of our proposed decentralized algorithms with algorithms originally devised for related problems, after applying them to our setting. Our comparison is in terms of sample complexity and Nash regret, and includes Independent Projected Policy Gradient (IPPG) (Leonardos et al. 2022), Independent Policy Gradient Ascent (IPGA) (Ding et al. 2022), Frank-Wolfe with Exploration I (FW-EI) (Cui et al. 2022), and Bandit Gradient Descent with Caratheodory Exploration (BGD-CE) (Dadi et al. 2024). Here, $n \in \mathbb{N}$ is the number of agents, $T \in \mathbb{N}$ is the number of rounds and $\varepsilon \geq 0$ is the desired tolerance for an ε -approximate Nash stable partition. For each algorithm, we specify the original problem for which it has been developed and whether it is projected-based.

Algorithm	Sample Complexity	Nash Regret	Original Problem	Projection?
IPPG (Leonardos et al. 2022)	$O(n^n/\varepsilon^6)$	NA	Potential Games	Yes
IPGA (Ding et al. 2022)	$O(n^n/\varepsilon^5)$	$O(n^n T^{4/5})$	Potential Games	Yes
FW-EI (Cui et al. 2022)	$O(n^{24}/\varepsilon^6)$	$O(n^4 T^{5/6})$	Congestion Games	No
BGD-CE (Dadi et al. 2024)	$O(n^{26.5}/\varepsilon^5)$	$O(n^{2.8} T^{4/5})$	Congestion Games	No
D1S-FW (ours)	$O(n^8/\varepsilon^2)$	$O(n^4 T^{1/2})$	Hedonic Games	No
DMS-FW (ours)	$O(n^{24}/\varepsilon^6)$	$O(n^4 T^{5/6})$	Hedonic Games	No

between 1 to n coalitions). In our project teams example, this can be thought of as if each developer picks which room to enter among n rooms (see Remark 2 for elaborations on this modelling choice). Thus, each agent i joins a certain coalition among n candidate ones at time t following a ***mixed strategy*** φ_i^t , built based only on *local* information from past rounds. Time $t = 1$ is exceptional, where each agent i initializes her strategy φ_i^1 to any *full support* strategy (e.g., the uniform strategy $\varphi_i^1(x) = \frac{1}{n}$), a mild condition we use in Section 4.1 to guarantee that the least-squares gradient estimator is well-defined. Formally, $\varphi_i^t \in \mathcal{S}_n$ where $\mathcal{S}_n$ is the probability simplex over $[n]$, i.e., for any $x \in [n]$, agent i picks the x th candidate coalition with probability $\varphi_i^t(x) \in [0, 1]$. Let $\varphi^t = (\varphi_i^t)_{i \in N} \in \mathcal{S}_n^n$ be the agents' ***joint mixed strategy*** at time t . At any time t , each agent i then samples an ***assignment*** $x_i \in [n]$ from φ_i^t independently from other agents, forming a ***joint assignment*** $\mathbf{x} = (x_i)_{i \in N}$. Thus, the agents' iterative process of distributed decision-making at time t unfolds as follows:

- (1) Each agent i samples $x_i \sim \varphi_i^t$ and joins the x_i th candidate coalition, forming a joint assignment $\mathbf{x} = (x_i)_{i \in N}$.
- (2) Each agent i observes the other members in the coalition she joined and gets bandit feedback about the utility gained from her own coalition.
- (3) Based *only* on this obtained information, agent i updates her strategy to be φ_i^{t+1} and moves to the next round $t + 1$.

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

Afterwards, we can derive each agent's utility from her chosen strategy, determined by aggregating her utilities of other agents. We focus on ***additively separable hedonic games*** (ASHGs) with *symmetric* preferences, where any pair of agents assign the same numerical value toward each other, indicating the intensity by which they prefer each other to another agent. As customary in the literature (see, e.g., (Flammini, Kodric, et al. 2021)), we assume that agents' valuations are within $[-1, 1]$. Recall that preferences are *unknown* and even the agents themselves may not be aware of them. Thus, for any pair of distinct agents i, j , the uncertainty about their mutual valuation

is captured by an *unknown* and *fixed* distribution $\mathcal{D}_{i,j}$ over $[-1, 1]$ with mean $d_{i,j}$, which agents i, j aim to learn. At any time t , the **utility** $v_{i,j}^t$ of agents i, j for each other is then independently drawn from $\mathcal{D}_{i,j}$. We use the convention that $v_{i,i}^t = d_{i,i} = 0$ for any agent i . For any joint assignment $\mathbf{x}$ sampled at time t , agent i 's utility from the induced partition $\pi^{\mathbf{x}}$ is then $v_i^t(\mathbf{x}) = \sum_{j \in \pi^{\mathbf{x}}(i)} v_{i,j}^t$, whose mean is $d_i(\mathbf{x}) = \sum_{j \in \pi^{\mathbf{x}}(i)} d_{i,j}$. Agent i 's utility from her strategy ϕ_i^t at time t is thus defined as $V_i(\boldsymbol{\phi}^t) := \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\phi}^t} [d_i(\mathbf{x})]$.

Each agent's strategies are evaluated by her *true* unknown utility, but she cannot observe her *true* utility. We thus consider the most general **decentralized** setting where each agent should learn her own preferences from repeated *partial* feedbacks. At any time t and for any joint assignment $\mathbf{x}$, we assume that each agent i can only receive **bandit feedback**, i.e., in practice, agent i only obtains her utility from the entire partition induced by $\mathbf{x}$ (i.e., $v_i^t(\mathbf{x})$), with no information about the utility $v_{i,j}^t$ gained from interacting with any other agent j .

Each agent joins a coalition with the goal of maximizing her own utility. We thus want to study stability under single agents' incentives to deviate between coalitions. The traditional literature on hedonic games focuses on **pure** strategies (e.g., (Aziz, Brandt, et al. 2013; Bilò et al. 2018)), where each agent i 's strategy is joining a *single* candidate coalition at time t by only selecting some $x_i \in [n]$. Let $\mathbf{x}_{-i} = (x_j)_{j \neq i}$ be the joint strategy of all agents except for agent i . Agent i can then **deviate** by moving from her selected coalition to another one with index $y_i \in [n]$, which is a **Nash deviation** if it improves her utility, i.e., $V_i^t(\mathbf{x}_{-i}, y_i) > V_i^t(\mathbf{x}_{-i}, x_i)$. Unlike prior work on hedonic games, we also study *mixed* strategies, where we define the other notion of *mixed Nash deviations*. Consider the joint strategy $\boldsymbol{\phi}^t$ at time t . Letting $\boldsymbol{\phi}_{-i}^t = (\phi_j^t)_{j \neq i}$ for any agent i , agent i may perform a (**mixed single-agent**) **deviation** from her strategy ϕ_i^t to another strategy $\phi_i \in \mathcal{S}_n$, which is a **mixed Nash deviation** only if it immediately makes her better off, i.e., $V_i(\boldsymbol{\phi}_{-i}^t, \phi_i) > V_i(\boldsymbol{\phi}^t)$. Hence, a (*pure* or *mixed*) joint strategy for which no Nash deviation is possible is said to be **Nash stable** (NS), also called a *Nash equilibrium* (NE). As mentioned earlier, for *symmetric* preferences under *pure* strategies, the existence of a Nash stable outcome is guaranteed by potential function argument (Bogomolnaia and Jackson 2002), but computing such strategies is PLS-complete (Gairing and Savani 2019).

We thus further consider an *approximate* notion of Nash stability. At each time t , note that agent i 's *best response* to the other agents' strategies is a Nash deviation given by a strategy $\phi_i^{*,t}$ satisfying $V_i^*(\boldsymbol{\phi}_{-i}^t) := V_i(\boldsymbol{\phi}_{-i}^t, \phi_i^{*,t}) = \max_{\phi \in \mathcal{S}_n} V_i(\boldsymbol{\phi}_{-i}^t, \phi)$. Thus, for any $\varepsilon \geq 0$, the agents' joint strategy $\boldsymbol{\phi}^t$ at time t is **ε -approximate Nash stable** (ε -NS) if no agent can improve her gain by more than ε , i.e., $\max_{i \in N} (V_i^*(\boldsymbol{\phi}_{-i}^t) - V_i(\boldsymbol{\phi}^t)) \leq \varepsilon$. Here, the quantity $\max_{i \in N} (V_i^*(\boldsymbol{\phi}_{-i}^t) - V_i(\boldsymbol{\phi}^t))$ measures the **worst agent's local gap** between the expected utilities she receives from her best response and her current strategy at time t .

REMARK 1 (LEARNING MEANINGFUL MIXED NASH STABLE STRATEGIES). Consider the mixed strategy where each agent treats all coalitions as equally desirable, i.e., the mixed strategy ϕ_i^t of any agent i and each time t satisfies $\phi_i^t(x) = \frac{1}{n}$ for any $x \in [n]$. Clearly, this is an exact mixed NS strategy. However, 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 meaningful mixed NS strategies that align with agents' preferences, reflecting real-life behavior by accounting for agents' incentives.

Our goal is thus devising a **decentralized** algorithm that learns an ε -NS joint strategy for some $\varepsilon \geq 0$, which aligns with agents' preferences. Putting everything together, at any time t , each agent i updates her strategy from time $t - 1$ based only on her own *local* history up to time $t - 1$. This history consists only of her strategies, the composition of the coalitions she joined and the bandit feedbacks from her own utilities until time $t - 1$, without any information about other coalitions and utilities. We analyze the performance of such algorithm via **Nash regret** (Ding et al. 2022; Q. Liu et al. 2021), comparing each agent's learnt strategy with her best response strategy

at any time instant. Formally, given a sequence of joint strategies $\{\boldsymbol{\varphi}^t\}_{t=1}^T$, the Nash regret after T rounds is:

$$\mathcal{R}^T := \sum_{t=1}^T \max_{i \in N} (V_i^*(\boldsymbol{\varphi}_{-i}^t) - V_i(\boldsymbol{\varphi}^t)). \quad (1)$$

The intuition behind the Nash regret lies in our definition of approximate NS strategies. As $\max_{i \in N} (V_i^*(\boldsymbol{\varphi}_{-i}^t) - V_i(\boldsymbol{\varphi}^t))$ is the worst agent's local gap at time t , the Nash regret in (1) is the total sum of the worst agent's local gap at each individual round. That is, at each time t , the Nash regret compares learned strategy φ_i^t of each agent i with the best strategy agent i can take by fixing the other agents' strategies φ_{-i}^t . By our definition of an approximate NS strategy, the Nash regret thus evaluates how far the agents' strategies in each round are from being an (approximate) NS strategy.

For any $\varepsilon \geq 0$, it is well-known that the above connection between ε -NS strategies and Nash regret can be used to show that an algorithm with a Nash regret bound of ε obtains an ε -NS strategy (see, e.g., (Ding et al. 2022)). Hence, one of our objectives is minimizing the Nash regret, i.e., attaining a **Nash regret bound** that is sublinear in the number of rounds T and polynomial in the number of agents n . Another goal is finding an ε -NS strategy using a number of rounds T that is *small* in its dependency on the number of agents n and $1/\varepsilon$, guaranteeing a **(PAC) sample complexity bound**. An algorithm with low sample complexity requires a fewer rounds for best-iterate convergence to an ε -NS strategy. In fact, any algorithm with sublinear Nash regret can be directly converted to a polynomial-sample algorithm via standard online-to-batch conversion (see, e.g., (Jin et al. 2018)).

REMARK 2 (MODELLING CHOICES). *We model coalition formation as a decentralized process in which each agent simultaneously selects one of n candidate coalitions. The candidate coalitions are merely labels indexing the possible coalitions of a partition and do not represent a centralized coordinator. Indeed, by Remark 3, each agent only uses the identities of the members of her own coalition, i.e., $[\psi_i(\mathbf{x}^t)]_j = \mathbb{1}\{j \in \pi^t(i)\}$, without any agent or third party learning global preferences or computing a partition. Our model differs from two common approaches in the literature. Unlike sequential-arrival models (Bullinger and Romen 2025b; Cohen 2026a; Cohen and Agmon 2024a; Flammini, Monaco, et al. 2021), where a central authority irrevocably assigns agents as they arrive, we consider simultaneous decisions made without coordination. Likewise, unlike deviation-dynamics models (Bilò et al. 2018; Boehmer et al. 2023; Brandt, Bullinger, and Tappe 2022), where agents repeatedly deviate under known preferences, we assume that agents initially do not know their own preferences and must learn them online from bandit feedback.*

Consequently, an agent's decision within a round cannot depend on the realized choices of other agents, as this would require intra-round communication that our decentralized setting explicitly excludes. Strategic interaction instead arises through the expected utilities $V_i(\boldsymbol{\varphi}^t) = \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\varphi}^t} [d_i(\mathbf{x})]$, which depend on the joint strategy of all agents and render each round's game a potential game (as proven later in Lemma 3.1).

3.1 DOL-ASHGs as Series of Potential Games

We begin with a useful structural property satisfied by our model. For a single round (i.e., $T = 1$), a Nash stable strategy is guaranteed for ASHG with *known* symmetric preferences through a potential function argument (Bogomolnaia and Jackson 2002). In general DOL-ASHGs, where the symmetric preferences are *unknown* at any time t , we show that the game associated with each time t is also a potential game:

LEMMA 3.1 (DOL-ASHGs ARE A SEQUENCE OF POTENTIAL GAMES). *At each time t , the ASHG with symmetric and unknown preferences associated with time t is a potential game.*

PROOF. Consider a joint mixed strategy $\boldsymbol{\varphi}^t$ at time t . Given a joint assignment $\mathbf{x} \sim \boldsymbol{\varphi}^t$, note that $\Phi^t(\mathbf{x}) = \frac{1}{2} \sum_{i \in N} v_i^t(\mathbf{x})$ is a potential function for *pure* strategies as $\Phi^t(\mathbf{x}_{-i}, x_i) - \Phi^t(\mathbf{x}_{-i}, x'_i) = v_i^t(\mathbf{x}_{-i}, x_i) - v_i^t(\mathbf{x}_{-i}, x'_i)$ for any agent i and another assignment $x'_i \in [n]$ of agent i . Therefore, by slight abuse of notation, $\Phi(\boldsymbol{\varphi}^t) = \sum_{i \in N} V_i(\boldsymbol{\varphi}^t)$

is a potential function for *mixed* strategies as $\Phi(\varphi_{-i}^t, \varphi_i^t) - \Phi(\varphi_{-i}^t, \phi_i) = V_i(\varphi_{-i}^t, \varphi_i^t) - V_i(\varphi_{-i}^t, \phi_i)$ for any agent i and another strategy $\phi_i \in \mathcal{S}_n$ of agent i . Note that, for pure strategies, the potential is denoted as Φ^t as it directly depends on t through the utilities. Yet, for mixed strategies, it is expressed by $\Phi(\varphi^t)$ as it does not directly depend on t , while its input φ^t does. $\square$

Thus, online learning algorithms for our context should be based on the principle that any **Nash stable** strategy is a stationary point of the potential function from Lemma 3.1. Other key design factors are computational efficiency and scalability. Note that one main challenge in our setting is that the size of the joint pure strategies' space equals to the number of possible partitions over n agents which grows *exponentially* with the number of agents n (Sandholm et al. 1999). Thus, an efficient algorithm should exhibit Nash regret and sample complexity polynomial in the number of agents n , without dependence on the size of the joint pure strategies space.

As mentioned in Section 2, existing methods for online learning in potential games are inappropriate for our context (Cui et al. 2022; Ding et al. 2022; Leonardos et al. 2022). Hence, we devise two Frank-Wolfe methods (Hazan and Kale 2012), which replace projections with a less expensive linear optimization step, and thus can be implemented efficiently in high-dimensional domains such as ours (See Section 3.2 for an overview of the classical Frank-Wolfe algorithm). Tailored to decentralized online learning in ASHG, our algorithms also leverage each agent's knowledge about the identities of other members in the coalitions she joins. Namely, each agent makes decisions based only on her own history of strategies, the members of the coalitions she joined and the resulting utility feedbacks, without communicating with other agents to coordinate their strategies. Instead of using large mini-batches as done by Cui et al. (Cui et al. 2022), in our first algorithm each agent only uses *one* sample per round (Section 4), yet it achieves Nash regret and sample complexity bounds that are **optimal** up to logarithmic factors. At each round, agents form coalitions, obtain a *single* utility feedback, update their strategies and move to the next round (i.e., agents consecutively follow the stages (1), (2), (3) from the iterative process specified in Section 3). For reducing the error suffered at each single time step, we then devise a method using *multiple* samples per round (Section 5), albeit resulting in *suboptimal* Nash regret and sample complexity bounds. Agents obtain multiple samples by splitting the rounds into several windows, where during each window agents repeatedly form coalitions and receive utility feedbacks over multiple steps following the same strategy. The agents update their strategies only at the end of each window. Namely, regarding the iterative process specified in Section 3, agents repeat the stages (1) and (2) multiple times across each window before proceeding to stage (3).

3.2 Overview of the Classical Frank-Wolfe Algorithm

The classic Frank-Wolfe algorithm can be used to solve constrained convex optimization problems of the form²:

$$\max_{x \in \mathcal{P}} f(x), \quad (2)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex, differentiable real-valued function and $\mathcal{P} \subseteq \mathbb{R}^n$ is some convex feasible region.

Instead of relying on potentially expensive and intractable projection operations used by traditional projection-based methods (see, e.g., (Nesterov et al. 2018; Nocedal and Wright 2006)), the Frank-Wolfe algorithm uses a linear minimization oracle, which is a subroutine that only involves minimizing a linear function over the feasible region, often proven to be more cost-effective than common projection-based approaches (see, e.g., (Combettes and Pokutta 2021) for a comprehensive comparison). Formally, the Frank-Wolfe algorithm is presented in Algorithm 1. At each iteration t , it approximates the objective function using its first-order Taylor expansion around the current iterate and solves the resulting linear problem over the feasible set (line 3). Specifically, the algorithm first determines a feasible descent direction by optimizing this linear approximation (line 3). It then selects an

²Although the standard Frank-Wolfe algorithm is typically presented for constrained convex *minimization* problems, we adopt the analogous formulation for constrained convex *maximization* problems, as they are the primary focus of our work and this perspective provides greater clarity for the ensuing discussion.

Algorithm 1 The Classical Frank-Wolfe Algorithm**Input:** A convex and differentiable real-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$; A convex feasible region $\mathcal{P} \subseteq \mathbb{R}^n$; T rounds.

- 1: Initialize an arbitrary initial point $\mathbf{x}_0 \in \mathcal{P}$.
- 2: **for each** time $t = 0, \dots, T - 1$ **do**
- 3: **Compute a feasible direction:** $\mathbf{v}_t \leftarrow \arg \max_{v \in \mathcal{P}} \langle \nabla f(\mathbf{x}_t), v \rangle$.
- 4: **Choose a step size:** Set $\gamma_t \leftarrow \frac{2}{t+2}$.
- 5: **Update:** $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \gamma_t (\mathbf{v}_t - \mathbf{x}_t)$.

appropriate step size (line 4), whose choice typically guides the convergence guarantees of the algorithm. For instance, Jaggi (Jaggi 2013) picks a step size of $\gamma_t = \frac{2}{t+2}$ at each iteration t . Finally, the algorithm uses the chosen step size to update the current solution using a convex combination of the previous iterate and the new direction (line 5).

4 Distributed One-Sample Algorithm

In this section, we devise a sample-efficient decentralized online learning algorithm for DOL-ASHGs (Algorithm 2), termed as *Decentralized One-Sample Frank-Wolfe* (D1S-FW), that requires only *one* sample per round to update the optimization variables and obtains sublinear Nash regret that is also polynomial in the number of agents n . As input, D1S-FW receives the number of rounds T , a set of n agents and step sizes $\rho_t, \eta_t, \mu_t \in (0, 1)$ for each time t whose choices affect the resulting Nash regret and sample complexity bounds, as detailed later. Next, we depict D1S-FW's main components. Note that we do not have access to the *exact* gradient of the potential function from Lemma 3.1. Thus, at any time t , each agent i first computes an unbiased and variance reduced gradient estimation of the potential function at time t using only *one* utility feedback (Section 4.1). Using this estimation, agent i then follows the classic Frank-Wolfe update rule to pick her next strategy (Section 4.2). Finally, we prove that D1S-FW has Nash regret and sample complexity bounds that are *optimal* up to logarithmic factors (Section 4.3).

4.1 The Gradient Estimation Step

Our algorithm first modifies the momentum variance reduction approach in (Mokhtari et al. 2018, 2020; M. Zhang et al. 2020) to compute an unbiased estimator of the potential's gradient using a *single* utility feedback. Formally, at each time t , note that the potential's gradient w.r.t. any joint strategy φ is given by $\nabla \Phi = (\frac{\partial \Phi}{\partial \varphi_i}(\mathbf{x}_i))_{i \in N, \mathbf{x}_i \in [n]}$, or equivalently $\nabla \Phi = (\nabla_{\varphi_i} \Phi)_{i \in N}$ where $\nabla_i \Phi := \nabla_{\varphi_i} \Phi = (\frac{\partial \Phi}{\partial \varphi_i}(\mathbf{x}_i))_{\mathbf{x}_i \in [n]}$ is the potential's gradient w.r.t. agent i 's strategy φ_i . A naive approach for agent i is thus using an unbiased *one-sample* estimation $\widehat{\nabla}_i^t \Phi$ at time t of $\nabla_i \Phi(\varphi^t)$. However, this method yields a high variance. To reduce the variance, in our algorithm agent i sets her gradient estimation $\mathbf{g}_i^t$ at time t to be a weighted average involving the previous gradient estimate $\mathbf{g}_i^{t-1}$ and the one-sample estimation $\widehat{\nabla}_i^t \Phi$:

$$\mathbf{g}_i^t = (1 - \mu_t) (\mathbf{g}_i^{t-1} + \widehat{\Delta}_i^t) + \mu_t \widehat{\nabla}_i^t \Phi, \quad (3)$$

where $\mu_t \in (0, 1)$ is an averaging parameter and $\widehat{\Delta}_i^t$ is an unbiased estimator of the gradient variation $\Delta_i^t = \nabla_i \Phi(\varphi^t) - \nabla_i \Phi(\varphi^{t-1})$. Intuitively, adding the term $\widehat{\Delta}_i^t$ to $\mathbf{g}_i^{t-1}$ in (3) ensures that $\mathbf{g}_i^t$ is an unbiased estimator of $\nabla_i \Phi$ due to an inductive argument. Indeed, if $\mathbf{g}_i^{t-1}$ is an unbiased estimate of $\nabla_i \Phi(\varphi^{t-1})$ and $\widehat{\Delta}_i^t$ is an unbiased estimate of Δ_i^t , then $\mathbb{E}[\mathbf{g}_i^t] = (1 - \mu_t)(\nabla_i \Phi(\varphi^{t-1}) + \nabla_i \Phi(\varphi^t) - \nabla_i \Phi(\varphi^{t-1})) + \mu_t \nabla_i \Phi = \nabla_i \Phi$, where the expectation is over all randomness up to time t . Next, we explain how to compute the one-sample gradient estimator $\widehat{\nabla}_i^t \Phi$ and the gradient variation estimator $\widehat{\Delta}_i^t$, both in an unbiased manner. Particularly, we prove that this can be done by approximating each agent's mean utilities from others using linear regression.

Algorithm 2 Decentralized One-Sample Frank-Wolfe (D1S-FW)

Input: T rounds; n agents; Step sizes $\rho_t, \eta_t, \mu_t \in (0, 1) \forall t$.

- 1: Initialize a *full-support* initial policy φ_i^1 for each agent i (e.g., the uniform strategy $\varphi_i^1(x) = \frac{1}{n}$).
- 2: **for each** time $t = 1, \dots, T$ **do**
- 3: Each agent i joins a coalition by sampling $x_i^t \sim \varphi_i^t$.
- 4: Each agent i receives a utility v_i^t from her coalition.
- 5: **for each** agent $i \in N$ individually **do**
- 6: Compute $\widehat{\nabla}_i^t \Phi$ using (8).
- 7: **if** $t = 1$ **then**
- 8: Set the gradient estimation as $\mathbf{g}_i^1 := \widehat{\nabla}_i^1 \Phi$.
- 9: **else**
- 10: Compute $\widehat{\Delta}_i^t = \widehat{\nabla}_i^t \Phi - \widehat{\nabla}_i^{t-1} \Phi$.
- 11: Compute the gradient estimator $\mathbf{g}_i^t$ using (3).
- 12: Calculate $\hat{\phi}_i^t = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \mathbf{g}_i^t \rangle$.
- 13: Compute a G-optimal design π_i^t using (9) (See Remark 5).
- 14: Set $\varphi_i^{t+1} = (1 - \rho_t)[\varphi_i^t + \eta_t(\hat{\phi}_i^t - \varphi_i^t)] + \rho_t \pi_i^t$.

One-Sample Gradient Estimator. To come up with such estimator, we first need to know the *closed form of the true gradient*. In Lemma 4.1, we show that, for each agent i and any selection $x_i \in [n]$, the gradient $(\nabla_i \Phi)_{x_i} := \frac{\partial \Phi}{\partial \varphi_i}(x_i)$ evaluated at the joint strategy $\boldsymbol{\varphi}^t$ can be written as an inner product between the vector $\xi_i = [d_{i,j}]_{j \in N}$ of all agent i 's mean utilities from any agent $j \in N$ and the vector $\boldsymbol{\varphi}^t(x_i) \in [0, 1]^n$ given by $[\boldsymbol{\varphi}^t(x_i)]_j = \varphi_j^t(x_i)$ for any joint strategy $\boldsymbol{\varphi}^t$ at time t . Intuitively, this will allow us to obtain our one-sample gradient estimator by proving that linear regression can be applied to approximate ξ_i at time t via a least squares estimator.

LEMMA 4.1. *For each agent i , the gradient evaluated at $\boldsymbol{\varphi}^t$, which is given by $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$, can be rephrased as:*

$$(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} = \langle \boldsymbol{\varphi}^t(x_i), \xi_i \rangle, \quad (4)$$

where $\boldsymbol{\varphi}^t(x_i) \in [0, 1]^n$ is defined as $[\boldsymbol{\varphi}^t(x_i)]_j = \varphi_j^t(x_i)$ for any joint strategy $\boldsymbol{\varphi}^t$ at time t , $x_i \in [n]$ and agent $j \in N$.

PROOF. For any joint strategy $\boldsymbol{\varphi}$, the potential function is:

$$\Phi(\boldsymbol{\varphi}) = \sum_{i \in N} \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\varphi}} [d_i(\mathbf{x})] = \sum_{i \in N} \sum_{x_i \in [n]} \varphi_i(x_i) \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}} [d_i(\mathbf{x}_{-i}, x_i)]. \quad (5)$$

For any $x_i, x_j \in [n]$, let $\mathbb{1}\{x_j = x_i\}$ be equal to 1 if $x_j = x_i$ and 0 otherwise. Hence, as the gradient $\nabla_i \Phi := \nabla_{\varphi_i} \Phi$ is of dimension n for any agent i , the gradient's x_i th coordinate for any $x_i \in [n]$ is given by:

$$\begin{aligned} (\nabla_i \Phi)_{x_i} &:= \frac{\partial \Phi}{\partial \varphi_i}(x_i) = \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}} [d_i(\mathbf{x}_{-i}, x_i)] = \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}} \left[\sum_{j \in \pi^{(\mathbf{x}_{-i}, x_i)}(i)} d_{i,j} \right] \\ &= \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}} \left[\sum_{j \in N} \mathbb{1}\{x_j = x_i\} d_{i,j} \right] = \sum_{j \in N} \mathbb{E}_{\mathbf{x}_j \sim \boldsymbol{\varphi}_j} [\mathbb{1}\{x_j = x_i\} d_{i,j}] \\ &= \sum_{j \in N} \sum_{y \in [n]} \varphi_j(y) \mathbb{E}[\mathbb{1}\{x_j = x_i\} d_{i,j} | x_j = y] = \sum_{j \in N} \varphi_j(x_i) d_{i,j}, \end{aligned} \quad (6)$$

from which (4) readily follows. □

Using Lemma 4.1, at any time t , each agent i can derive a **one-sample gradient estimation** as follows. First, agent i samples her chosen candidate coalition $x_i^t \in [n]$ from her strategy φ_i^t (line 3). After forming the partition $\pi^t := \pi^{\mathbf{x}^t}$, each agent i obtains a utility feedback v_i^t from her coalition $\pi^t(i)$ (line 4), whose expectation given that the agents' joint assignment is $\mathbf{x}^t \in [n]^n$ at time t can be expressed as $\mathbb{E}[v_i^t | \mathbf{x}^t] = d_i(\mathbf{x}^t) = \sum_{i \neq j \in \pi^t(i)} d_{i,j} = \sum_{i \neq j \in N} \mathbb{1}\{x_j^t = x_i^t\} d_{i,j}$, where, for any $x_i, x_j \in [n]$, $\mathbb{1}\{x_j = x_i\}$ equals to 1 if $x_j = x_i$ and 0 otherwise. This can be rephrased as follows:

$$\mathbb{E}[v_i^t | \mathbf{x}^t] = \langle \boldsymbol{\psi}_i(\mathbf{x}^t), \boldsymbol{\xi}_i \rangle, \quad (7)$$

where $\boldsymbol{\psi}_i(\mathbf{x}^t) \in \{0, 1\}^n$ is given by $[\boldsymbol{\psi}_i(\mathbf{x}^t)]_j = \mathbb{1}\{x_j^t = x_i^t\}$ for any agent $j \in N$, capturing whether agent j joined agent i 's selected coalition or not.

REMARK 3 (COMPUTING $\boldsymbol{\psi}_i(\mathbf{x}^t)$). *In practice, each agent i can compute $\boldsymbol{\psi}_i(\mathbf{x}^t)$ by only using her knowledge about other agents in the coalition she joined via $[\boldsymbol{\psi}_i(\mathbf{x}^t)]_j = \mathbb{1}\{j \in \pi^t(i)\}$ for any agent j . From the perspective of agent i , she deduces $\boldsymbol{\psi}_i(\mathbf{x}^t)$ only from the coalition $\pi^t(i)$ she joined, but we include the implicit dependence on the agents' joint selection $\mathbf{x}^t$ for the sake of the analysis so as to, e.g., reason about the covariance matrix.*

Therefore, by Lemma 4.1, **we can use linear regression to estimate $\boldsymbol{\xi}_i$** at time t via the least squares estimator $\hat{\boldsymbol{\xi}}_i^t = (\hat{\Psi}_i^t)^+ \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t$, where $\hat{\Psi}_i^t = \mathbb{E}_{\mathbf{x}_i \sim \varphi_i^t} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top]$ is the covariance matrix and $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. We use the pseudoinverse rather than the ordinary inverse because $\hat{\Psi}_i^t$ is in general singular (see Lemma A.1 in Appendix A). Nonetheless, the estimator remains well-defined and *unbiased*. Indeed, note that every feature vector $\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)$ lies in the span $\mathcal{U}_i^t := \text{span}\{\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)\}_{x_i \in [n]}$. Moreover, agent i 's strategy φ_i^t is enforced to have full support, as ensured for time $t = 1$ by the full-support initialization in line 1, and for time $t \geq 2$ by the full-support exploratory design π_i^t (see Remark 5). Therefore, we have $\text{range}(\hat{\Psi}_i^t) = \mathcal{U}_i^t$, and the pseudoinverse acts as the ordinary inverse on $\mathcal{U}_i^t$. Consequently, all our guarantees are preserved verbatim (see Lemma A.2 in Appendix A). We can then obtain an estimator $\widehat{\nabla}_i^t \Phi(x_i) = \langle \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i), \hat{\boldsymbol{\xi}}_i^t \rangle$ of $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$, which satisfies the following:

$$\widehat{\nabla}_i^t \Phi(x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^+ \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t. \quad (8)$$

One can easily verify that this is indeed an *unbiased* estimator of the potential's gradient, as we prove in Appendix A.

One-Sample Gradient Variation Estimator. For any agent i , our one-sample approach for estimating the gradient variation $\Delta_i^t = \nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})$ at time t is computing the difference between our one-sample gradient estimation for agent i 's current selection and that for her previous one, i.e., $\widehat{\Delta}_i^t = \widehat{\nabla}_i^t \Phi - \widehat{\nabla}_i^{t-1} \Phi$. As we saw earlier (recall Lemma A.3), $\widehat{\nabla}_i^t \Phi$ and $\widehat{\nabla}_i^{t-1} \Phi$ are unbiased estimators of $\nabla_i \Phi$ at times t and $t - 1$ (respectively), and thus $\mathbb{E}[\widehat{\Delta}_i^t] = \Delta_i^t$, where the expectation is over all randomness up to time t . That is, $\widehat{\Delta}_i^t$ is an *unbiased* estimator of the gradient variation Δ_i^t .

REMARK 4 (INAPPLICABILITY OF TWO-POINT ZERO-ORDER ESTIMATORS). *One may ask why we do not reduce variance using two-point (or multi-point) zeroth-order gradient estimators (Duchi et al. 2015), which achieve low, dimension-robust variance by evaluating the objective at two learner-chosen points under a common realization of the randomness. Such methods are inapplicable in our setting for two reasons. First, agent i 's feedback $v_i^t(\mathbf{x}) = \sum_{i \neq j \in \pi^t(i)} v_{i,j}^t$ depends on the joint assignment $\mathbf{x} = (x_i, \mathbf{x}_{-i})$. Since agent i controls only x_i , she cannot evaluate the objective at prescribed perturbations, as the remaining coordinates $\mathbf{x}_{-i}$ (and thus the coalition she joins) are chosen by the other agents. Second, each round produces a single bandit feedback under freshly sampled, independent utilities $v_{i,j}^t \sim \mathcal{D}_{i,j}$ (line 4), preventing the paired-query noise cancellation that underlies two-point methods. We therefore estimate $\nabla_i \Phi$ from a single feedback via linear regression on the closed form $(\nabla_i \Phi)_{x_i} = \langle \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i), \boldsymbol{\xi}_i \rangle$ (Lemma 4.1), and reduce variance temporally through the momentum term in (3).*

4.2 The Frank-Wolfe Update Step

At each time t , in line 12 each agent i then optimizes a linear optimization problem to get the feasible ascent strategy $\hat{\phi}_i^t$ as the regular Frank-Wolfe method (i.e., as in line 3 of Algorithm 1). Instead of updating her strategy at window $t + 1$ to be $\phi_i^t + \eta_t(\hat{\phi}_i^t - \phi_i^t)$ using the standard Frank-Wolfe update (i.e., as in line 5 of Algorithm 1), in line 14 agent i mixes this strategy with an exploratory strategy π_i^t which is computed in line 13. Specifically, agent i computes a *G-optimal design* π_i^t for $\{\psi_i(\mathbf{x}_{-i}^t, x_i)\}_{x_i \in [n]}$ (see, e.g., (Lattimore and Szepesvári 2020, Chapter 21)). The G-optimal design minimizes the maximal mean-squared prediction error in linear regression. In our context, the G-optimal design can be viewed as the strategy allowing a *uniformly well* estimation of ξ_i over all candidate coalitions. That is, sampling a candidate coalition $x_i \in [n]$ in proportion to $\pi_i^t(x_i)$ minimizes the number of samples required for reaching a desired level of accuracy in estimating $\mathbb{E}[v_i^t | \mathbf{x}_{-i}^t, x_i] = \langle \psi_i(\mathbf{x}_{-i}^t, x_i), \xi_i \rangle$. Formally, agent i 's **G-optimal design** π_i^t at time t satisfies:

$$\pi_i^t \in \arg \min_{\phi \in \mathcal{S}_n} f_i^t(\phi), \quad (9)$$

where $f_i^t(\phi) = \max_{x_i \in [n]} \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i,\phi}^t)^{-1} \psi_i(\mathbf{x}_{-i}^t, x_i)$. Hence, agent i then mixes π_i^t with the strategy $\phi_i^t + \eta_t(\hat{\phi}_i^t - \phi_i^t)$ obtained from the standard Frank-Wolfe update.

REMARK 5 (COMPUTING APPROXIMATE G-OPTIMAL DESIGNS). While the original G-optimal design problem is NP-hard (Soare et al. 2014), for our purposes it suffices to compute an approximate G-optimal design with minimal impact on performance. For instance, using the Frank-Wolfe algorithm and a proper initialization, agent i can efficiently find an approximate G-optimal design $\hat{\pi}_i^t$ with $f_i^t(\hat{\pi}_i^t) \leq 2n$. For easing the analysis in the main text, we hereafter assume that a G-optimal design can be found efficiently and accurately, which is an assumption common in the literature (see, e.g., (Yang and Tan 2022)). See Appendix J for a vast discussion on the minor effect of using an approximate G-optimal design in our algorithm, including a time complexity analysis. We further take the exploratory design π_i^t to have full support (e.g., by mixing the approximate G-optimal design with a $\frac{1}{n}$ -uniform floor, which affects the design's value by at most a constant factor). By line 14, this makes ϕ_i^t full-support for every $t \geq 2$, and thus $\text{range}(\hat{\Psi}_i^t) = \mathcal{U}_i^t$. Together with the full-support initialization for $t = 1$, this suffices for Lemma A.2 and thus for the well-definedness and unbiasedness of the estimator (8).

4.3 Nash Regret Analysis of D1S-FW

In this section, we prove that D1S-FW (Algorithm 2) has Nash regret and sample complexity bounds that are **optimal** up to logarithmic factors. To obtain the Nash regret bound of D1S-FW, the key step is devising an upper bound on our estimators' errors. First, mixing the regular Frank-Wolfe update with the exploration strategy π_i^t allows us to regulate the one-sample gradient estimator's error as follows:

LEMMA 4.2. At each time t , for all $\delta \in (0, 1]$, agent i and $x_i \in [n]$, the following holds with probability at least $1 - \delta$:

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} \right| \leq O \left(\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (10)$$

In particular, if $\rho_t \leq \frac{2 \log(1/\delta)}{9}$, then the following holds with probability at least $1 - \delta$:

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} \right| \leq O \left(\frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (11)$$

PROOF. (Sketch) At any time t , let $\mathcal{R}_t = \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i,\phi_i^t}^t)^{-1} \psi_i(\mathbf{x}_{-i}^t, x_i) v_i^t$. In Appendix C, we first prove that $|\mathcal{R}_t| \leq \frac{n^2}{\rho_t}$ (Lemma C.1) and $\mathbb{E}^t[\mathcal{R}_t^2] \leq \frac{n^3}{\rho_t}$ (Lemma C.2), where $\mathbb{E}^t[\cdot]$ is the expectation over $\mathbf{x}_i^t \sim \varphi^t$ for each agent

i and all the utilities' randomness at time t . As we proved in Lemma A.3 that $\widehat{\nabla}_i^t \Phi(x_i)$ from (8) is an unbiased estimator of $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$ (i.e., $\mathbb{E}^t[\widehat{\nabla}_i^t \Phi(x_i)] = \mathbb{E}[\mathcal{R}_t] = (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$), we can apply Bernstein's inequality to obtain both (10) and (11). $\square$

We next supply the required upper bound on the error $\mathcal{A}_i^t := \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty$ of agent i 's gradient estimator in (3), where $\|\cdot\|_\infty$ is the L_∞ norm (i.e., $\|\mathbf{z}\|_\infty = \max_{i \in [n]} |z_i|$ for any $\mathbf{z} \in \mathbb{R}^n$).

LEMMA 4.3. *For any time $t \geq 3$ and any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for any agent i :*

$$\|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty \leq O\left(\sum_{\tau_1=2}^{t-1} \mathcal{B}^{\tau_1} \prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) + \mathcal{B}^t\right), \quad (12)$$

where $\mathcal{B}^t := \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t}$. If $\rho_t \leq \frac{2 \log(1/\delta)}{9}$ for any time t , then an improved bound in (12) is obtained for $\mathcal{B}^t := \frac{n^3 \log(nT/\delta)}{\rho_t}$. For $t \in \{1, 2\}$, the errors $\mathcal{A}_i^1$ and $\mathcal{A}_i^2$ are bounded as in Lemma 4.2 since $\mathcal{A}_i^t = \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi\|_\infty$ for $t \in \{1, 2\}$.

PROOF. (Sketch) As $\mathbf{g}_i^1 = \widehat{\nabla}_i^1 \Phi$, then $\mathcal{A}_i^1$ is bounded as in Lemma 4.2. For any time $t \geq 2$, we first prove in Appendix D that our estimators satisfy $\mathcal{A}_i^t := \|(1 - \mu_t)(\nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1}) + (1 - \mu_t)(\widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})) + (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi)\|_\infty$ for each agent i , from which our bound for time 2 can be easily obtained. For each time $t \geq 3$, this equality allows us to prove (12) by induction on t . $\square$

We are now ready to establish our Nash regret bound:

THEOREM 4.1. *For any $\delta \in (0, 1]$, D1S-FW (Algorithm 2) with $\rho_t = \frac{1}{t}$, $\eta_t = \frac{1}{\sqrt{t}}$, $\mu_t = 1 - \frac{1}{t}$ for any time t and $T \geq 2$ obtains the following Nash regret bound with probability at least $1 - \delta$:*

$$\mathcal{R}^T \leq O\left((n + n^3 + n^4)\sqrt{T} + n^4 \log(nT/\delta)\right). \quad (13)$$

If $\delta \leq e^{-\frac{9}{2T}}$, then, with probability at least $1 - \delta$, the Nash regret bound in (13) can be reduced to:

$$\mathcal{R}^T \leq O\left((n + n^3 + n^4)\sqrt{T} + n^4 \log(nT/\delta)\right). \quad (14)$$

PROOF. (Sketch) The key idea behind our proof is using the fact that the Nash regret is at most the sum of the Frank-Wolfe gaps across all rounds (Appendix B). Namely, as the Frank-Wolfe gap of any joint strategy $\boldsymbol{\varphi}$ w.r.t. the potential function $\Phi(\boldsymbol{\varphi})$ is given by $G(\boldsymbol{\varphi}) = \max_{\boldsymbol{\varphi}' \in S_n^n} \langle \boldsymbol{\varphi}' - \boldsymbol{\varphi}, \nabla_{\boldsymbol{\varphi}} \Phi \rangle$, then $\mathcal{R}^T \leq \sum_{t \in [T]} G(\boldsymbol{\varphi}^t)$, and thus we derive an upper bound on the right-hand side to attain our Nash regret bound. First, consider the joint strategies $\hat{\boldsymbol{\varphi}}^t = (\hat{\varphi}_i^t)_{i \in N}$ and $\boldsymbol{\pi}^t = (\pi_i^t)_{i \in N}$ obtained from the strategies computed in lines 12-13. As the potential function $\Phi(\boldsymbol{\varphi}) = \sum_{i \in N} V_i(\boldsymbol{\varphi})$ is n^2 -smooth w.r.t. the L_1 -norm $\|\cdot\|_1$ by Lemma E.1 in Appendix E, then $\Phi(\boldsymbol{\varphi}^{t+1}) \geq \Phi(\boldsymbol{\varphi}^t) + \langle \nabla \Phi(\boldsymbol{\varphi}^t), \boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t \rangle - \frac{n^2}{2} \|\boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t\|_1^2$. Combined with line 14, we prove in Appendix E that:

$$\Phi(\boldsymbol{\varphi}^{t+1}) \geq \Phi(\boldsymbol{\varphi}^t) + (1 - \rho_t)\eta_t \langle \nabla \Phi(\boldsymbol{\varphi}^t), \hat{\boldsymbol{\varphi}}^t - \boldsymbol{\varphi}^t \rangle - \rho_t \cdot n^3 - \frac{n^4}{2} [(1 - \rho_t)^2 \eta_t^2 + \rho_t^2] \quad (15)$$

Thus, we then devise a lower bound on the right-hand side. In Appendix E, we first show that $\langle \nabla \Phi(\boldsymbol{\varphi}^t), \hat{\boldsymbol{\varphi}}^t - \boldsymbol{\varphi}^t \rangle \geq G(\boldsymbol{\varphi}^t) - 2n \|\nabla \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}^t\|_\infty$, where $\mathbf{g}^t = (\mathbf{g}_i^t)_{i \in [n]}$ is the gradient estimator of $\nabla \Phi(\boldsymbol{\varphi}^t)$. After applying this inequality to (15), we rearrange the resulting inequality to obtain:

$$G(\boldsymbol{\varphi}^t) \leq \frac{\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^t)}{(1 - \rho_t)\eta_t} + 2n \|\nabla \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}^t\|_\infty + \frac{\rho_t n^3}{(1 - \rho_t)\eta_t} + \frac{n^4(1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4}{2(1 - \rho_t)\eta_t}. \quad (16)$$

By summing this inequality over all rounds T while noting that $\sum_{t \in [T]} [\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^t)] = \Phi(\boldsymbol{\varphi}^{T+1}) - \Phi(\boldsymbol{\varphi}^1)$ as a telescoping series and $|\Phi(\boldsymbol{\varphi}^{T+1}) - \Phi(\boldsymbol{\varphi}^1)| \leq n$, we then have that the Nash regret is upper bounded by:

$$\mathcal{R}^T \leq 2n \sum_{t \in [T]} \|\nabla \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}^t\|_\infty + \frac{nT}{(1 - \rho_t)\eta_t} + \frac{\rho_t n^3 T}{(1 - \rho_t)\eta_t} + \frac{n^4 T(1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4 T}{2(1 - \rho_t)\eta_t}. \quad (17)$$

Recall that this bound also serves as an upper bound on the Nash regret incurred by our algorithm since $\mathcal{R}^T \leq \sum_{t \in [T]} G(\boldsymbol{\varphi}^t)$. Further, our bound in (17) depends on the algorithm's error term, for which we have already supplied an upper bound in Lemma 4.3. Hence, after plugging the algorithm's inputs $\rho_t = \frac{1}{T}$, $\eta_t = \frac{1}{\sqrt{T}}$, $\mu_t = 1 - \frac{1}{T}$ for any time t and $T \geq 2$ to (17), in Appendix E we then apply Lemma 4.3 to obtain the Nash regret bounds in (13)–(14). $\square$

4.3.1 Discussion: D1S-FW's Nash regret and sample complexity bounds. Theorem 4.1 indicates that D1S-FW has a Nash regret bound of $\tilde{O}(\sqrt{T})$. By standard online-to-batch conversion (see, e.g., (Jin et al. 2018)), our Nash regret bound suggests a sample complexity bound of $T = O(n^8/\epsilon^2)$ for obtaining an ϵ -NS strategy. Surprisingly, though D1S-FW uses only *one* sample per round, our Nash regret and sample complexity bounds are **optimal** up to logarithmic factors (Bai and Jin 2020; Hassani et al. 2020). To guarantee those bounds, Theorem 4.1 dictates that the step sizes should depend on the number of rounds T , which is known in our context. As mentioned earlier, this is practical in many realistic cases as our project teams example from Section 1, where developers may participate in project teams over a predefined number of project milestones. Even if T is not explicitly known in some cases, developers often have a rough estimate based on the nature of open-source projects. For example, they may anticipate a typical project duration based on prior experience or the project's roadmap. Hence, agents can still execute our algorithm in a distributed manner even without knowing T .

5 Distributed Multi-Sample Algorithm

While our *one-sample* approach in Section 4 is more practical and efficient than prior methods, as discussed in Section 4.3.1, using several utility feedbacks for gradient estimation allows us to better control the estimation variance. We thus propose a sample-efficient decentralized algorithm for DOL-ASHGs (Algorithm 3), termed *Decentralized Multi-Sample Frank-Wolfe* (DMS-FW), which uses *multiple* samples per round while maintaining sublinear Nash regret that is polynomial in the number of agents n .

Specifically, we split the number of rounds T into $M > 0$ windows, each with $K > 0$ steps. At each window $m \in [M]$, let $\boldsymbol{\varphi}^m = (\varphi_i^m)_{i \in N}$ denote the agents' joint mixed strategy, where $\varphi_i^m \in \mathcal{S}_n$ is agent i 's strategy over the n candidate coalitions. Throughout the K steps of window m , each agent i follows the same strategy φ_i^m and collects K utility feedbacks, which she then uses to construct our multi-sample gradient estimator (Section 5.1). Agent i subsequently uses this estimator to update her strategy for window $m + 1$ (Section 5.2). In terms of the iterative process introduced in Section 3, each agent therefore repeats stages (1) and (2) for K steps within each window before proceeding to stage (3). Finally, we show that increasing K reduces the gradient estimation error at each round, but comes at the cost of *suboptimal* Nash regret and sample complexity bounds (Section 5.3).

5.1 The Gradient Estimation Step

Our multi-sample gradient estimator is similar to our one-sample estimation from Section 4.1. Formally, each agent i treats the number of rounds T as being composed of $M > 0$ windows, each of size $K > 0$ (i.e., $T = MK$), where K and M are given as input to our algorithm. At each window $m \in [M]$, agent i plays according to the same strategy φ_i^m across the K steps, i.e., during any step $k \in [K]$ of the window m , agent i first samples her chosen candidate coalition $x_i^{m,k} \in [n]$ from φ_i^m (line 4). After forming the partition $\pi^{m,k} := \pi^{x^{m,k}}$, agent i obtains

Algorithm 3 Decentralized Multi-Sample Frank-Wolfe (DMS-FW)**Input:** $T = MK$ rounds; n agents; step sizes $\rho, \eta \in (0, 1)$.

- 1: Initialize an arbitrary initial policy φ_i^1 for each agent i .
- 2: **for each** window $m = 1, \dots, M$ **do**
- 3: **for each** step $k = 1, \dots, K$ **do**
- 4: Each agent i joins a coalition by sampling $x_i^{m,k} \sim \varphi_i^m$.
- 5: Each agent i receives a utility $v_i^{m,k}$ from her assignment.
- 6: **for each** agent $i \in N$ individually **do**
- 7: For each $x_i \in [n]$, calculate $\widehat{\nabla}_i^{m,K} \Phi(x_i)$ using (18).
- 8: Calculate $\hat{\phi}_i^m = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \widehat{\nabla}_i^{m,K} \Phi \rangle$.
- 9: Compute $\{\pi_i^{m,k}\}_{k \in [K]}$ using (19) (by arguments similar to Remark 5).
- 10: Set the exploratory strategy as $\pi_i^m = \frac{1}{K} \sum_{k \in [K]} \pi_i^{m,k}$.
- 11: Set $\varphi_i^{m+1} = (1 - \rho)[\varphi_i^m + \eta(\hat{\phi}_i^m - \varphi_i^m)] + \rho\pi_i^m$.

the utility feedback $v_i^{m,k}$ (line 5). As in (7), $v_i^{m,k}$'s expectation given the agents' joint assignment $\mathbf{x}^{m,k} \in [n]^n$ satisfies $\mathbb{E}[v_i^{m,k} | \mathbf{x}^{m,k}] = \langle \psi_i(\mathbf{x}^{m,k}), \xi_i \rangle$.

Recalling Remark 3, we can similarly employ linear regression to estimate ξ_i at step k of window m using $\hat{\xi}_i^{m,k} = (\Psi_{i,\varphi_i^m}^{m,k})^+ \psi_i(\mathbf{x}^{m,k}) v_i^{m,k}$, where $\Psi_{i,\phi}^{m,k} = \mathbb{E}_{x_i \sim \phi} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top]$ for any $\phi \in \mathcal{S}_n$ and $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. As for the one-sample estimator, the covariance matrix $\Psi_{i,\varphi_i^m}^{m,k}$ is generally singular (Lemma A.1), while the pseudoinverse yields a well-defined and unbiased estimator under the same conditions (Lemma A.2). Thus, the analysis of Lemma 5.1 is remains unchanged. Recalling that the gradient $\nabla_i \Phi := \nabla_{\varphi_i} \Phi$ is of dimension n for any agent i , $\widehat{\nabla}_i^{m,K} \Phi(x_i) = \frac{1}{K} \sum_{k \in [K]} \langle \psi_i(\mathbf{x}_{-i}^{m,k}, x_i), \hat{\xi}_i^{m,k} \rangle$ is thus an estimator of the gradient's x_i th coordinate $(\nabla_i \Phi(\varphi^m))_{x_i} := \frac{\partial \Phi(\varphi^m)}{\partial \varphi_i}(x_i)$ for any $x_i \in [n]$, where that estimator equals to:

$$\widehat{\nabla}_i^{m,K} \Phi(x_i) = \frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\varphi_i^m}^{m,k})^+ \psi_i(\mathbf{x}_{-i}^{m,k}, x_i) v_i^{m,k}. \quad (18)$$

In Appendix F, we prove that $\widehat{\nabla}^{m,K} \Phi := (\widehat{\nabla}_i^{m,K} \Phi(x_i))_{i \in N, x_i \in [n]}$ is indeed an *unbiased* estimator of the potential's gradient.

5.2 The Frank-Wolfe Update Step

At each window m , in line 8 each agent i then optimizes a linear optimization problem to get the feasible ascent strategy $\hat{\phi}_i^m$ following the regular Frank-Wolfe method (i.e., as in line 3 of Algorithm 1). Afterwards, in line 11 agent i mixes this strategy with an exploratory strategy π_i^m which is computed in lines 9–10 as follows. As in Section 4.2, for any step k , agent i computes a G-optimal design $\pi_i^{m,k}$ for $\{\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)\}_{x_i \in [n]}$ via:

$$\pi_i^{m,k} \in \arg \min_{\phi \in \mathcal{S}_n} g^{m,k}(\phi) \quad (19)$$

where $g^{m,k}(\phi) = \max_{x_i \in [n]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\phi}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)$. Thus, agent i obtains her exploratory strategy π_i^m in line 10 by averaging the G-optimal designs in line 9, which she then mixes with the strategy attained from the standard Frank-Wolfe update.

5.3 Nash Regret Analysis of DMS-FW

To establish the Nash regret bound of DMS-FW (Algorithm 3), the crucial idea is deriving an upper bound on the error incurred by our gradient estimator in (18). Similarly to Lemma 4.2, this error is bounded as follows:

LEMMA 5.1. *At any window m , for all $\delta \in (0, 1]$, agent i and $x_i \in [n]$, the following holds with probability at least $1 - \delta$:*

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) - (\nabla_i \Phi(\varphi^m))_{x_i} \right| \leq O \left(\sqrt{\frac{n^4 \log(nM/\delta)}{\rho K}} + \frac{n^3 \log(nM/\delta)}{\rho K} \right). \quad (20)$$

In particular, if $\rho \leq \frac{2 \log(1/\delta)}{9}$, then, with probability at least $1 - \delta$:

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) - (\nabla_i \Phi(\varphi^m))_{x_i} \right| \leq O \left(\frac{4n^3 \log(nM/\delta)}{\rho K} \right). \quad (21)$$

PROOF. (Sketch) The proof follows from arguments similar to Lemma 4.2 and is thus deferred to Appendix G. $\square$

We are now ready to derive our Nash regret bound:

THEOREM 5.1. *Consider $T = MK$ rounds and let $\delta \in (0, 1]$. We distinguish two regimes.*

- **General regime:** DMS-FW (Algorithm 3) with $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $M \geq 2$ and $K = M^2$ (so that $T = M^3$) obtains the following Nash regret bound with probability at least $1 - \delta$:

$$\mathcal{R}^T \leq O \left(\left(n + 2n^3 + n^4 + \sqrt{n^6 \mathcal{F}_\delta^T} \right) T^{5/6} + n^4 T^{2/3} \left(1 + \mathcal{F}_\delta^T \right) \right), \quad (22)$$

where $\mathcal{F}_\delta^T := \log(nT^{1/3}/\delta)$. Moreover, the choice $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $K = M^2$ is asymptotically optimal for the general bound: no choice of the window/step split $T = MK$ and step sizes $\rho, \eta \in (0, 1)$ makes the upper bound proved for DMS-FW smaller than $\tilde{O}(T^{5/6})$ (see Lemma I.2 in Appendix I).

- **Reduced regime:** If $\delta \leq e^{-\frac{9}{2M}}$, then the split $K = M^2$ is no longer optimal; the optimal choice is $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $M \geq 2$ and $K = M^{3/2}$ (so that $T = M^{5/2}$, i.e., $M = T^{2/5}$), under which, with probability at least $1 - \delta$, the Nash regret bound improves to:

$$\mathcal{R}^T \leq O \left((n + n^3 + n^4) T^{4/5} + n^4 \cdot T^{4/5} \mathcal{G}_\delta^T \right), \quad (23)$$

where $\mathcal{G}_\delta^T := \log(nT^{2/5}/\delta)$. This choice is asymptotically optimal for the reduced bound: no choice of $T = MK$ and $\rho, \eta \in (0, 1)$ makes the (reduced) upper bound proved for DMS-FW smaller than $\tilde{O}(T^{4/5})$ (see Lemma I.3 in Appendix I).

PROOF. (Sketch) As in Theorem 4.1, since the Nash regret is at most the sum of the Frank-Wolfe gaps across all rounds (Appendix B), in Appendix H we devise an upper bound on the latter that is similar to (17) and relies on the estimator's error. Concretely, using $\mathcal{R}^T \leq K \sum_{m \in [M]} G(\varphi^m)$ together with the per-window smoothness argument (identical to the proof of Theorem 4.1 in Appendix E, applied per window and with the per-round factor K), we obtain the parameter-free bound (142) of Appendix I. Applying Lemma 5.1 and substituting $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $K = M^2$ yields (22), while, in the reduced regime, substituting $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $K = M^{3/2}$ yields (23). Finally, minimizing (142) over all splits $T = MK$ and step sizes ρ, η establishes the asymptotic optimality of these choices (Lemmas I.2 and I.3 in Appendix I). $\square$

5.3.1 Discussion: The Stability-Accuracy Tradeoff. We thus observe an interesting tradeoff between an algorithm's estimation accuracy³ during each individual round and its overall convergence guarantees. By Lemmas 4.2 and 5.1, DMS-FW's error at each time t can be made lower than that of D1S-FW (Algorithm 2) by increasing the number of steps K . However, by Theorems 4.1 and 5.1, this comes at the expense of approximate Nash stability (and ultimately the convergence rate). Specifically, Theorem 5.1 indicates that DMS-FW has a Nash regret bound of $\tilde{O}(T^{5/6})$ in the general regime, translating (as in Section 4.3.1) to a sample complexity bound of $O(n^{24}/\epsilon^6)$ for obtaining an ϵ -NS strategy. In the reduced (high-probability) regime, the *optimal* split $K = M^{3/2}$ of Theorem 5.1 improves the Nash regret bound to $\tilde{O}(T^{4/5})$ and the sample complexity to $\tilde{O}(n^{20}/\epsilon^5)$. Either way, D1S-FW (Algorithm 2) has Nash regret and sample complexity bounds that are **optimal** up to logarithmic factors, while DMS-FW (Algorithm 3) obtains only suboptimal bounds.

We stress that DMS-FW is *not* a strict generalization of D1S-FW. Hence, the regret bound of DMS-FW does not reduce to the $\tilde{O}(\sqrt{T})$ regret bound of D1S-FW as $M \rightarrow 1$. The two algorithms reduce variance by different mechanisms: D1S-FW uses *temporal* momentum via the parameter μ_t in (3), whereas DMS-FW uses *in-window averaging* over K samples (see (18)) and carries no momentum term. Consequently, no choice of (M, K) makes DMS-FW coincide with D1S-FW; in particular, setting $K = 1$ removes all variance reduction and yields a vacuous (superlinear) bound, as we show in Appendix I (Proposition I.1). The gap between the two algorithms is thus an inherent consequence of the different variance-reduction strategies, rather than an artifact of the analysis.

6 Conclusions and Future Work

In this paper, we presented a new algorithmic framework for studying online learning in coalition formation from the viewpoint of *decentralized decision-making* by selfish agents. Our goal was designing sample-efficient distributed algorithms for self-interested agents that balance long- and short-term performance, trading off between accuracy at each round and the long-term objective of minimizing Nash regret to reach approximately Nash stable outcomes. As such, we devised two Frank-Wolfe methods. In our first algorithm, every agent sacrifices accuracy at each round by only using *one* sample per round for gradient estimation, yet it attains Nash regret and sample complexity bounds that are **optimal** up to logarithmic factors. Our second method reduces the error suffered at each single round by employing *multiple* samples per round, but yields *suboptimal* Nash regret and sample complexity bounds.

Our work opens the way for plenty of future studies. An immediate direction includes the investigation of other classes of hedonic games, such as fractional hedonic games (Aziz, Brandl, et al. 2019), modified fractional hedonic games (Monaco et al. 2020) and social distance games (Balliu et al. 2019). Another interesting direction is investigating other solution concepts, like popularity (Brandt and Bullinger 2022). Future work also warrants exploring other models of partial information (e.g., semi-bandit or noisy feedback), studying cases where the number of rounds is not necessarily known. Future work could also investigate dynamic agent populations in which the set of participating agents evolves over time through arrivals and departures. Another intriguing future direction is devising distributed algorithms with no-regret guarantees for any agent adopting them, ensuring that each agent has diminishing regret, regardless of how others update their strategies. Finally, this work concentrated on laying the theoretical foundations for decentralized online learning in coalition formation, which we plan to evaluate empirically as part of our ongoing and future work. An interesting direction is to evaluate our framework empirically, potentially drawing on experimental platforms for interactive coalition formation such as the Online Coalition Game (Wissink et al. 2022).

³Here, by *accuracy* we mean the accuracy of an agent's gradient estimator at a single round (Lemmas 4.2 and 5.1), *not* the social-welfare optimality of the resulting partition; the term "stability" refers to (approximate) Nash stability as measured by the Nash regret in (22). Social welfare enters our analysis only as the potential function $\Phi(\varphi) = \sum_{i \in N} V_i(\varphi)$ (Lemma 3.1) and is not itself an evaluation metric.

Acknowledgments

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

References

- A. Aloisio, M. Flammini, and C. Vinci. 2020. “The impact of selfishness in hypergraph hedonic games.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 2, 1766–1773.
- H. Aziz, F. Brandl, F. Brandt, P. Harrenstein, M. Olsen, and D. Peters. 2019. “Fractional hedonic games.” *ACM Transactions on Economics and Computation (TEAC)*, 7, 2, 1–29.
- H. Aziz, F. Brandt, and H. G. Seedig. 2013. “Computing desirable partitions in additively separable hedonic games.” *Artificial Intelligence*, 195, 316–334.
- H. Aziz and R. Savani. 2016. “Hedonic Games.” In: *Handbook of Computational Social Choice*. Ed. by F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. Cambridge University Press, 356–376.
- Y. Bai and C. Jin. 2020. “Provable Self-Play Algorithms for Competitive Reinforcement Learning.” In: *International Conference on Machine Learning*. Vol. 119. PMLR, 551–560.
- M.-F. Balcan, A. D. Procaccia, and Y. Zick. 2015. “Learning cooperative games.” In: *Proceedings of the 24th International Conference on Artificial Intelligence*, 475–481.
- C. Ballester. 2004. “NP-completeness in hedonic games.” *Games and Economic Behavior*, 49, 1, 1–30.
- A. Balliu, M. Flammini, G. Melideo, and D. Olivetti. 2019. “On non-cooperativeness in social distance games.” *Journal of Artificial Intelligence Research*, 66, 625–653.
- S. Banerjee, H. Konishi, and T. Sönmez. 2001. “Core in a simple coalition formation game.” *Social Choice and Welfare*, 18, 1, 135–153.
- V. Bilò, A. Fanelli, M. Flammini, G. Monaco, and L. Moscardelli. 2018. “Nash stable outcomes in fractional hedonic games: Existence, efficiency and computation.” *Journal of Artificial Intelligence Research*, 62, 315–371.
- F. Bloch and E. Diamantoudi. 2011. “Noncooperative formation of coalitions in hedonic games.” *International Journal of Game Theory*, 40, 263–280.
- N. Boehmer, M. Bullinger, and A. M. Kerkmann. 2023. “Causes of stability in dynamic coalition formation.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 5, 5499–5506.
- A. Bogomolnaia and M. O. Jackson. 2002. “The stability of hedonic coalition structures.” *Games and Economic Behavior*, 38, 2, 201–230.
- E. Boursier and V. Perchet. 2024. “A Survey on Multi-player Bandits.” *Journal of Machine Learning Research*, 25, 137, 1–45.
- F. Brandt and M. Bullinger. 2022. “Finding and recognizing popular coalition structures.” *Journal of Artificial Intelligence Research*, 74, 569–626.
- F. Brandt, M. Bullinger, and L. Tappe. 2022. “Single-agent dynamics in additively separable hedonic games.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 5, 4867–4874.
- M. Bullinger and R. Romen. 2025a. “Online coalition formation under random arrival or coalition dissolution.” *ACM Transactions on Algorithms*, 22, 1, 1–43.
- M. Bullinger and R. Romen. 2025b. “Stability in online coalition formation.” *Journal of Artificial Intelligence Research*, 82, 2423–2452.
- M. Bullinger, R. Romen, and A. Schlenga. 2026. “The Power of Matching for Online Fractional Hedonic Games.” In: *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2163–2194.
- X. Chen and B. Peng. 2020. “Hedging in games: Faster convergence of external and swap regrets.” *Advances in Neural Information Processing Systems*, 33, 18990–18999.
- S. Cohen. 2026a. “Delayed Assignments in Online Non-Centroid Clustering with Stochastic Arrivals.” In: *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS-26*, 1091–1100.
- S. Cohen. 2026b. “Offline Learning of Nash Stable Coalition Structures with Possibly Overlapping Coalitions.” In: *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS-26*, 3722–3730.
- S. Cohen and N. Agmon. 2023a. “Complexity of probabilistic inference in random dichotomous hedonic games.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 5, 5573–5581.
- S. Cohen and N. Agmon. 2025a. “Decentralized Online Learning by Selfish Agents in Coalition Formation.” In: *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*.
- S. Cohen and N. Agmon. 2023b. “Online Coalitional Skill Formation.” In: *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 494–503.
- S. Cohen and N. Agmon. 2024a. “Online Friends Partitioning Under Uncertainty.” In: *ECAI 2024*. IOS Press, 3332–3339.
- S. Cohen and N. Agmon. 2025b. “Online Learning of Coalition Structures by Selfish Agents.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 39, 13, 13709–13717.
- S. Cohen and N. Agmon. 2025c. “Online Learning of Fair Coalition Structures.” In: *Proceedings of the 28th European Conference on Artificial Intelligence, ECAI-25*, 4937–4944.

- S. Cohen and N. Agmon. 2024b. "Online Learning of Partitions in Additively Separable Hedonic Games." In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2722–2730.
- C. W. Combettes and S. Pokutta. 2021. "Complexity of linear minimization and projection on some sets." *Operations Research Letters*, 49, 4, 565–571.
- Q. Cui, Z. Xiong, M. Fazel, and S. S. Du. 2022. "Learning in congestion games with bandit feedback." *Advances in Neural Information Processing Systems*, 35, 11009–11022.
- L. Dadi, I. Panageas, S. Skoulakis, L. Viano, and V. Cevher. 2024. "Polynomial Convergence of Bandit No-Regret Dynamics in Congestion Games." *arXiv preprint arXiv:2401.09628*.
- C. Daskalakis, M. Fishelson, and N. Golowich. 2021. "Near-optimal no-regret learning in general games." *Advances in Neural Information Processing Systems*, 34, 27604–27616.
- C. Daskalakis and Q. Pan. 2014. "A counter-example to Karlin's strong conjecture for fictitious play." In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 11–20.
- A. Defazio and L. Bottou. 2019. "On the ineffectiveness of variance reduced optimization for deep learning." In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1755–1765.
- D. Ding, C.-Y. Wei, K. Zhang, and M. Jovanovic. 2022. "Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence." In: *International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 162, 5166–5220.
- M. Drechsler. 2025. "Learning coalition formation under an agglomeration bonus: Impacts on coalition structure and scheme performance." *Resource and Energy Economics*, 83, 101512.
- J. H. Dr  ze and J. Greenberg. 1980. "Hedonic coalitions: Optimality and stability." *Econometrica: Journal of the Econometric Society*, 987–1003.
- J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. 2015. "Optimal rates for zero-order convex optimization: The power of two function evaluations." *IEEE Transactions on Information Theory*, 61, 5, 2788–2806.
- M. Feldman, L. Lewin-Eytan, and J. Naor. 2015. "Hedonic clustering games." *ACM Transactions on Parallel Computing (TOPC)*, 2, 1, 1–48.
- S. Fioravanti, M. Flammini, B. Kodric, and G. Varricchio. 2023. "PAC learning and stabilizing Hedonic Games: towards a unifying approach." *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 5, 5641–5648.
- M. Flammini, B. Kodric, G. Monaco, and Q. Zhang. 2021. "Strategyproof mechanisms for additively separable and fractional hedonic games." *Journal of Artificial Intelligence Research*, 70, 1253–1279.
- M. Flammini, G. Monaco, L. Moscardelli, M. Shalom, and S. Zaks. 2021. "On the Online Coalition Structure Generation Problem." *Journal of Artificial Intelligence Research*, 72, 1215–1250.
- M. Gairing and R. Savani. 2019. "Computing stable outcomes in symmetric additively separable hedonic games." *Mathematics of Operations Research*, 44, 3, 1101–1121.
- H. Hassani, A. Karbasi, A. Mokhtari, and Z. Shen. 2020. "Stochastic conditional gradient++: (non)convex minimization and continuous submodular maximization." *SIAM Journal on Optimization*, 30, 4, 3315–3344.
- E. Hazan and S. Kale. 2012. "Projection-free online learning." In: *International Conference on Machine Learning*, 521–528.
- A. Igarashi, J. Sliwinski, and Y. Zick. 2019. "Forming probably stable communities with limited interactions." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 01. Vol. 33, 2053–2060.
- M. Jaggi. 2013. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization." *International Conference on Machine Learning*, 28, 1, 427–435.
- C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. 2018. "Is Q-learning provably efficient?" *Advances in Neural Information Processing Systems*, 31, 4863–4873.
- T. Lattimore, C. Szepesvari, and G. Weisz. 2020. "Learning with good feature representations in bandits and in rl with a generative model." In: *International conference on machine learning*. PMLR, 5662–5670.
- T. Lattimore and C. Szepesv  ri. 2020. *Bandit algorithms*. Cambridge University Press.
- S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. 2022. "Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games." In: *International Conference on Learning Representations*.
- D. S. Leslie and E. J. Collins. 2006. "Generalised weakened fictitious play." *Games and Economic Behavior*, 56, 2, 285–298.
- L. T. Liu, H. Mania, and M. Jordan. 2020. "Competing bandits in matching markets." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 1618–1628.
- Q. Liu, T. Yu, Y. Bai, and C. Jin. 2021. "A Sharp Analysis of Model-based Reinforcement Learning with Self-Play." In: *International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 139, 7001–7010.
- C. Maheshwari, S. Sastry, and E. Mazumdar. 2022. "Decentralized, communication- and coordination-free learning in structured matching markets." *Advances in Neural Information Processing Systems*, 35, 15081–15092.
- A. Mokhtari, H. Hassani, and A. Karbasi. 2018. "Conditional gradient method for stochastic submodular maximization: Closing the gap." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 1886–1895.

- A. Mokhtari, H. Hassani, and A. Karbasi. 2020. “Stochastic conditional gradient methods: From convex minimization to submodular maximization.” *Journal of machine learning research*, 21, 105, 1–49.
- G. Monaco, L. Moscardelli, and Y. Velaj. 2020. “Stable outcomes in modified fractional hedonic games.” *Autonomous Agents and Multi-Agent Systems*, 34, 1, 1–29.
- Y. Nesterov et al.. 2018. *Lectures on convex optimization*. Vol. 137. Springer.
- J. Nocedal and S. Wright. 2006. *Numerical optimization*. Vol. 2. Springer.
- T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohmé. 1999. “Coalition structure generation with worst case guarantees.” *Artificial intelligence*, 111, 1-2, 209–238.
- J. Sliwinski and Y. Zick. 2017. “Learning Hedonic Games.” In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI-17*, 2730–2736.
- M. Soare, A. Lazaric, and R. Munos. 2014. “Best-arm identification in linear bandits.” *Advances in Neural Information Processing Systems*, 27.
- S.-C. Sung and D. Dimitrov. 2010. “Computational complexity in additive hedonic games.” *European Journal of Operational Research*, 203, 3, 635–639.
- M. J. Todd. 2016. *Minimum-volume ellipsoids: Theory and algorithms*. Vol. 23. SIAM.
- J. Wissink, I. Van Beest, T. Pronk, and N. van de Ven. 2022. “The online coalition game: A tool for online interactive coalition formation research.” *Behavior Research Methods*, 54, 3, 1078–1091.
- M. Xu and D. Klabjan. Mar. 2025. “Multi-agent Multi-armed Bandit Regret Complexity and Optimality.” In: *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*. Vol. 258. PMLR, (Mar. 2025), 1639–1647.
- J. Yang and V. Tan. 2022. “Minimax optimal fixed-budget best arm identification in linear bandits.” *Advances in Neural Information Processing Systems*, 35, 12253–12266.
- M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi. 2020. “One sample stochastic frank-wolfe.” In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 4012–4023.
- Y. Zhang, S. Wang, and Z. Fang. 2022. “Matching in Multi-arm Bandit with Collision.” *Advances in Neural Information Processing Systems*, 35, 9552–9563.

A The One-Sample Gradient Estimator is Unbiased

In this section, we first show that the covariance matrix $\hat{\Psi}_i^t$ is in general *singular*, so that the Moore–Penrose pseudoinverse $(\hat{\Psi}_i^t)^+$ used in (8) is necessary. We then prove that, under our full-support exploration, this pseudoinverse acts as a genuine inverse on the feature span, so that every guarantee of the paper is preserved. Finally, we prove that our one-sample estimation is indeed unbiased.

Throughout, fix a time t , an agent i and the realized $\mathbf{x}_{-i}^t$, and abbreviate $\boldsymbol{\psi}^{(\ell)} := \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, \ell) \in \{0, 1\}^n$ for $\ell \in [n]$, so that $[\boldsymbol{\psi}^{(\ell)}]_j = \mathbb{1}\{x_j^t = \ell\}$ for $j \neq i$ and $[\boldsymbol{\psi}^{(\ell)}]_i = 1$. Recall $\hat{\Psi}_i^t = \Psi_{i, \varphi_i^t}^t = \sum_{\ell \in [n]} \varphi_i^t(\ell) \boldsymbol{\psi}^{(\ell)} \boldsymbol{\psi}^{(\ell)\top}$, and let $\mathcal{U}_i^t := \text{span}\{\boldsymbol{\psi}^{(\ell)} : \ell \in [n]\}$ denote the feature span.

LEMMA A.1 (SINGULARITY OF THE COVARIANCE MATRIX). *For any strategy φ_i^t , $\text{range}(\hat{\Psi}_i^t) = \text{span}\{\boldsymbol{\psi}^{(\ell)} : \varphi_i^t(\ell) > 0\} \subseteq \mathcal{U}_i^t$, and*

$$\dim \mathcal{U}_i^t = 1 + |\{\ell \in [n] : \exists j \neq i, x_j^t = \ell\}| \leq n. \quad (24)$$

In particular, if two distinct agents $j, k \neq i$ share a coalition, i.e., $x_j^t = x_k^t$, then $\mathbf{e}_j - \mathbf{e}_k \in \ker(\hat{\Psi}_i^t)$, so $\hat{\Psi}_i^t$ is singular. Hence, $\hat{\Psi}_i^t$ is invertible only in the non-generic event that the other $n - 1$ agents occupy pairwise distinct coalitions (and φ_i^t has full support).

PROOF. For $j \neq i$, agent j contributes $[\boldsymbol{\psi}^{(\ell)}]_j = 1$ for the unique $\ell = x_j^t$ and 0 otherwise, while $[\boldsymbol{\psi}^{(\ell)}]_i = 1$ for all ℓ . Thus $\boldsymbol{\psi}^{(\ell)} = \mathbf{e}_i + \mathbf{u}_\ell$, where $\mathbf{u}_\ell := \sum_{j \neq i: x_j^t = \ell} \mathbf{e}_j$. The vectors $\{\mathbf{u}_\ell\}_\ell$ have pairwise disjoint supports contained in $\{\mathbf{e}_j : j \neq i\}$, so the nonzero ones are linearly independent, and $\mathbf{e}_i \notin \text{span}\{\mathbf{u}_\ell\}$ (its support is $\{i\}$). Therefore $\dim \mathcal{U}_i^t = 1 + |\{\ell : \mathbf{u}_\ell \neq \mathbf{0}\}| = 1 + |\{\ell : \exists j \neq i, x_j^t = \ell\}|$, which proves (24); the bound $\leq n$ follows since the $n - 1$ other agents cannot occupy all n candidate coalitions. Since $\hat{\Psi}_i^t = \sum_{\ell} \varphi_i^t(\ell) \boldsymbol{\psi}^{(\ell)} \boldsymbol{\psi}^{(\ell)\top}$ is a nonnegative combination of rank-one terms, $\text{range}(\hat{\Psi}_i^t) = \text{span}\{\boldsymbol{\psi}^{(\ell)} : \varphi_i^t(\ell) > 0\}$. Finally, if $x_j^t = x_k^t =: \ell_0$ for distinct $j, k \neq i$, set $\mathbf{z} = \mathbf{e}_j - \mathbf{e}_k$.

For every ℓ , $\boldsymbol{\psi}^{(\ell)\top} \mathbf{z} = \mathbb{1}\{x_j^t = \ell\} - \mathbb{1}\{x_k^t = \ell\} = \mathbb{1}\{\ell = \ell_0\} - \mathbb{1}\{\ell = \ell_0\} = 0$, so $\mathbf{z}^\top \hat{\Psi}_i^t \mathbf{z} = \sum_\ell \varphi_i^t(\ell) (\boldsymbol{\psi}^{(\ell)\top} \mathbf{z})^2 = 0$; as $\hat{\Psi}_i^t \succeq 0$, this yields $\hat{\Psi}_i^t \mathbf{z} = \mathbf{0}$, so $\mathbf{z} \neq \mathbf{0}$ lies in $\ker(\hat{\Psi}_i^t)$ and $\hat{\Psi}_i^t$ is singular. $\square$

LEMMA A.2 (THE PSEUDOINVERSE IS EXACT ON THE FEATURE SPAN). *If φ_i^t has full support, then $\text{range}(\hat{\Psi}_i^t) = \mathcal{U}_i^t$, and, letting $P := \hat{\Psi}_i^t(\hat{\Psi}_i^t)^+ = (\hat{\Psi}_i^t)^+\hat{\Psi}_i^t$ be the orthogonal projector onto $\mathcal{U}_i^t$, we have $P\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)$ for every $x_i \in [n]$.*

PROOF. Full support gives $\text{range}(\hat{\Psi}_i^t) = \text{span}\{\boldsymbol{\psi}^{(\ell)} : \varphi_i^t(\ell) > 0\} = \text{span}\{\boldsymbol{\psi}^{(\ell)} : \ell \in [n]\} = \mathcal{U}_i^t$ by Lemma A.1. For a symmetric positive-semidefinite matrix, $\hat{\Psi}_i^t(\hat{\Psi}_i^t)^+ = (\hat{\Psi}_i^t)^+\hat{\Psi}_i^t = P$ is the orthogonal projector onto $\text{range}(\hat{\Psi}_i^t) = \mathcal{U}_i^t$. Since $\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) = \boldsymbol{\psi}^{(x_i)} \in \mathcal{U}_i^t$ for every $x_i \in [n]$, we get $P\boldsymbol{\psi}^{(x_i)} = \boldsymbol{\psi}^{(x_i)}$. $\square$

We are now ready to prove that our one-sample estimation is indeed unbiased.

LEMMA A.3. *At any time $t \in [T]$, the one-sample estimation $\hat{\nabla}_i^t \Phi(\mathbf{x})$ of the gradient $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$ is unbiased for each agent i and selection $x_i \in [n]$, i.e.:*

$$\mathbb{E}^t \left[\hat{\nabla}_i^t \Phi(x_i) \right] = (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}, \quad (25)$$

where $\mathbb{E}^t[\cdot]$ is the expectation over $x_i^t \sim \varphi^t$ for any agent i and all the utilities' randomness at time t .

PROOF. First, recall that our estimator is given by $\hat{\nabla}_i^t \Phi(x_i) = \langle \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i), \hat{\xi}_i^t \rangle$ of $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$, which satisfies the following:

$$\hat{\nabla}_i^t \Phi(x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^+ \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t. \quad (26)$$

Now, recall that the utility feedback v_i^t is agent i 's utility from the coalition she joined. As such, if $v_{i,j}^t$ is agent i 's utility from each agent $j \in N$, then $v_i^t = \sum_{j \in \pi^t(i)} v_{i,j}^t$. By the definition of $\boldsymbol{\psi}_i(\mathbf{x}^t)$, we infer that $v_i^t = \boldsymbol{\psi}_i(\mathbf{x}^t)^\top [v_{i,j}^t]_{j \in N}$. Therefore, given that $\mathbf{x}_{-i}^t$ is the joint selection of all agents other than agent i , after taking the expectation over $x_i^t \sim \varphi_i^t$ in (26), we have that:

$$\mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\hat{\nabla}_i^t \Phi(x_i) \right] = \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^+ \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t \right] = \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^+ \boldsymbol{\psi}_i(\mathbf{x}^t) \boldsymbol{\psi}_i(\mathbf{x}^t)^\top [v_{i,j}^t]_{j \in N} \right]. \quad (27)$$

Noting that only the quantity $\boldsymbol{\psi}_i(\mathbf{x}^t) \boldsymbol{\psi}_i(\mathbf{x}^t)^\top$ depends on $x_i^t \sim \varphi_i^t$, then:

$$\mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\hat{\nabla}_i^t \Phi(x_i) \right] = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^+ \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i^t) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i^t)^\top \right] [v_{i,j}^t]_{j \in N}. \quad (28)$$

Recalling that the covariance matrix is given by $\hat{\Psi}_i^t = \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i^t) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i^t)^\top \right]$ and that φ_i^t has full support, we invoke Lemma A.2: with $P := (\hat{\Psi}_i^t)^+ \hat{\Psi}_i^t$ the orthogonal projector onto $\mathcal{U}_i^t$, we have $(\hat{\Psi}_i^t)^+ \hat{\Psi}_i^t = P$ and, since P is symmetric and $P\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)$, it follows that $\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top P = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top$. Therefore:

$$\mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\hat{\nabla}_i^t \Phi(x_i) \right] = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top P [v_{i,j}^t]_{j \in N} = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top [v_{i,j}^t]_{j \in N}. \quad (29)$$

Now, recall that $\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \in \{0, 1\}^n$ is given by $[\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)]_j = \mathbb{1}\{x_j^t = x_i\}$ for any agent $j \in N$, capturing whether agent j joined agent i 's selected coalition or not. This means that $\mathbb{E}_{x_j^t \sim \varphi_j^t} [\mathbb{1}\{x_j^t = x_i\}] = \varphi_j^t(x_i)$. As such, by also taking the expectation over $\mathbf{x}_{-i}^t \sim \varphi_{-i}^t$, note that:

$$\mathbb{E}_{\mathbf{x}^t \sim \varphi^t} \left[\hat{\nabla}_i^t \Phi(x_i) \right] = \boldsymbol{\varphi}^t(x_i)^\top [v_{i,j}^t]_{j \in N}, \quad (30)$$

where we used our definition of $\boldsymbol{\varphi}^t(x_i) \in [0, 1]^n$ as $[\boldsymbol{\varphi}^t(x_i)]_j = \varphi_j^t(x_i)$ for any agent $j \in N$. Finally, by also taking the expectation over the utilities' randomness at time t , we conclude that:

$$\mathbb{E}^t \left[\hat{\nabla}_i^t \Phi(x_i) \right] = \boldsymbol{\varphi}^t(x_i)^\top \boldsymbol{\xi}_i = (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}, \quad (31)$$

where we used the fact that $(\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} = \langle \boldsymbol{\varphi}^t(x_i), \xi_i \rangle$ by equation (4). $\square$

B Relating Nash Stability and Nash Regret to the Frank-Wolfe Gap

In this section, we discuss how Nash stability and Nash regret relate to the Frank-Wolfe gap (Jaggi 2013), which is a well-known measure for defining stationarity that is defined as follows. For a joint strategy $\boldsymbol{\varphi}$, the Frank-Wolfe gap $G(\boldsymbol{\varphi})$ of $\boldsymbol{\varphi}$ with respect to the potential function is given by:

$$G(\boldsymbol{\varphi}) = \max_{\boldsymbol{\varphi}' \in S_n^n} \langle \boldsymbol{\varphi}' - \boldsymbol{\varphi}, \nabla_{\boldsymbol{\varphi}} \Phi \rangle. \quad (32)$$

Note that, for each agent i who fixes her strategy to be φ'_i , the following holds:

$$V_i(\boldsymbol{\varphi}_{-i}, \varphi'_i) - V_i(\boldsymbol{\varphi}) = \Phi(\boldsymbol{\varphi}_{-i}, \varphi'_i) - \Phi(\boldsymbol{\varphi}). \quad (33)$$

Recalling that in (6) we proved that $\nabla_{\varphi_i} \Phi(x_i) = \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}^t} [d_i(\mathbf{x}_{-i}, x_i)]$ for each agent i , then we infer from (33) that:

$$V_i(\boldsymbol{\varphi}_{-i}, \varphi'_i) - V_i(\boldsymbol{\varphi}) = \langle \varphi'_i - \varphi_i, \nabla_{\varphi_i} \Phi \rangle \leq \max_{\boldsymbol{\varphi}' \in S_n^n} \langle \boldsymbol{\varphi}' - \boldsymbol{\varphi}, \nabla_{\boldsymbol{\varphi}} \Phi \rangle = G(\boldsymbol{\varphi}). \quad (34)$$

Recall that, given a sequence of joint strategies $\{\boldsymbol{\varphi}^t\}_{t=1}^T$, the Nash regret after T rounds is given by $\mathcal{R}^T := \sum_{t=1}^T \max_{i \in N} (V_i^*(\boldsymbol{\varphi}_{-i}^t) - V_i(\boldsymbol{\varphi}^t))$. Consequently, due to (34), we infer that the Nash regret is upper bounded by the sum of the Frank-Wolfe gaps across all T rounds, that is:

$$\mathcal{R}^T \leq \sum_{t=1}^T G(\boldsymbol{\varphi}^t). \quad (35)$$

Moreover, recall that, for any $\varepsilon \geq 0$, the agents' joint strategy $\boldsymbol{\varphi}$ is ε -approximate Nash stable (ε -NS) if $\max_{i \in N} (V_i^*(\boldsymbol{\varphi}_{-i}) - V_i(\boldsymbol{\varphi})) \leq \varepsilon$. As such, we infer the following relation between the Frank-Wolfe gap and an approximately Nash-stable joint strategy:

$$\text{If } G(\boldsymbol{\varphi}) \leq \varepsilon, \text{ then } \boldsymbol{\varphi} \text{ is } \varepsilon\text{-NS by (34).} \quad (36)$$

C Proof of Lemma 4.2

Throughout this appendix, since φ_i^t has full support (Remark 5), every covariance matrix is invertible on the feature span $\mathcal{U}_i^t = \text{span}\{\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)\}_{x_i \in [n]}$ (Lemma A.1), and we write $(\cdot)^{-1}$ for the Moore–Penrose pseudoinverse $(\cdot)^+$, which restricted to $\mathcal{U}_i^t$ is a genuine inverse (Lemma A.2). As every vector appearing below lies in $\mathcal{U}_i^t$, all quadratic forms are well-defined. By the Kiefer–Wolfowitz theorem (Lattimore and Szepesvári 2020, Chapter 21), the effective dimension of the design is $d_i^t := \dim \mathcal{U}_i^t \leq n$, so the value n used below as an upper bound on $g(\pi_i^t)$ remains valid (and may be replaced by the smaller d_i^t).

Before presenting the proof of Lemma 4.2, we require two auxiliary results. The first lemma describes a universal upper bound on the one-sample gradient approximation $\widehat{\nabla}_i^t \Phi(x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t$ that is induced by the selected strategy at each time instant.

LEMMA C.1. *At each time t , the following holds for each agent i and candidate coalition $x_i \in [n]$:*

$$\left| \widehat{\nabla}_i^t \Phi(x_i) \right| = \left| \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t \right| \leq \frac{n^2}{\rho_t}. \quad (37)$$

PROOF. Recall that agent i 's strategy φ_i^t at time t is given as regular Frank-Wolfe update with the exploration strategy π_i^t , i.e., $\varphi_i^t = (1 - \rho_t)[\varphi_i^{t-1} + \eta_t(\phi_i^{t-1} - \varphi_i^{t-1})] + \rho_t \pi_i^t$. The covariance matrix at time t , which is given

by $\Psi_{i,\varphi_i^t}^t = \mathbb{E}_{x_i \sim \varphi_i^t} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top]$, therefore satisfies the following:

$$\begin{aligned} \Psi_{i,\varphi_i^t}^t &= \mathbb{E}_{x_i \sim \varphi_i^t} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top] \\ &= (1 - \rho_t) \mathbb{E}_{x_i \sim [\varphi_i^{t-1} + \eta(\phi_i^{t-1} - \varphi_i^{t-1})]} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top] + \rho_t \mathbb{E}_{x_i \sim \pi_i^{t-1}} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top] \\ &= (1 - \rho_t) \Psi_{i,\tilde{\varphi}_i^{t-1}}^t + \rho_t \Psi_{i,\pi_i^{t-1}}^t, \end{aligned} \quad (38)$$

where we denoted $\tilde{\varphi}_i^{t-1} := \varphi_i^{t-1} + \eta(\phi_i^{t-1} - \varphi_i^{t-1})$.

Further, recall that π_i^t is agent i 's G-optimal design for $\{\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)\}_{x_i \in [n]}$ during time t , which satisfies $\pi_i^t \in \arg \min_{\phi \in \mathcal{S}_n} g(\phi)$ for $g(\phi) = \max_{x_i \in [n]} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i,\phi}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)$. Accordingly, for each $\mathbf{z} \in \mathbb{R}^n$:

$$\mathbf{z}^\top \Psi_{i,\varphi_i^t}^t \mathbf{z} = (1 - \rho_t) \cdot \mathbf{z}^\top \Psi_{i,\tilde{\varphi}_i^{t-1}}^t \mathbf{z} + \rho_t \cdot \mathbf{z}^\top \Psi_{i,\pi_i^{t-1}}^t \mathbf{z} \geq \rho_t \cdot \mathbf{z}^\top \Psi_{i,\pi_i^{t-1}}^t \mathbf{z}. \quad (39)$$

We thus infer that:

$$\boldsymbol{\psi}_i(\mathbf{x}^t)^\top (\Psi_{i,\varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) \leq \frac{1}{\rho_t} \boldsymbol{\psi}_i(\mathbf{x}^t)^\top (\Psi_{i,\pi_i^{t-1}}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t). \quad (40)$$

By the Kiefer–Wolfowitz Theorem (Lattimore and Szepesvári 2020, Chapter 21), it is known that:

$$g(\pi_i^{t-1}) = \max_{x_i \in [n]} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i)^\top (\Psi_{i,\pi_i^{t-1}}^{t-1})^{-1} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i) = d_i^{t-1} \leq n. \quad (41)$$

In particular, note that:

$$\begin{aligned} n &\geq \max_{\mathbf{x}_{-i}^{t-1} \in [n]^{n-1}} g(\pi_i^{t-1}) \\ &= \max_{\mathbf{x}_{-i}^{t-1} \in [n]^{n-1}} \max_{x_i \in [n]} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i)^\top (\mathbb{E}_{x_i \sim \pi_i^{t-1}} [\boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i) \boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i)^\top])^{-1} \boldsymbol{\psi}_i(\mathbf{x}_{-i}^{t-1}, x_i) \\ &\geq \boldsymbol{\psi}_i(\mathbf{x}^t)^\top (\Psi_{i,\pi_i^{t-1}}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t). \end{aligned} \quad (42)$$

Applying (42) to (40) yields:

$$\boldsymbol{\psi}_i(\mathbf{x}^t)^\top (\Psi_{i,\varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) \leq \frac{n}{\rho_t}. \quad (43)$$

Note that this expression is essentially the norm of $\boldsymbol{\psi}_i(\mathbf{x}^t)$ with respect to the norm induced by the matrix $(\Psi_{i,\varphi_i^t}^t)^{-1}$. Specifically, for any matrix $A \in \mathbb{R}^{n \times n}$ and two vectors $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$, the inner product $\langle \cdot, \cdot \rangle_A$ induced by A is given as $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle_A := \mathbf{z}_1^\top A \mathbf{z}_2$, which reduces to the norm $\|\cdot\|_A$ induced by A via $\|\mathbf{z}_1\|_A^2 = \langle \mathbf{z}_1, \mathbf{z}_1 \rangle_A$. That is, (43) indicates that:

$$\|\boldsymbol{\psi}_i(\mathbf{x}^t)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}}^2 \leq \frac{n}{\rho_t} \iff \|\boldsymbol{\psi}_i(\mathbf{x}^t)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho_t}}. \quad (44)$$

By similar arguments, note that for any $x_i \in [n]$:

$$\|\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho_t}}. \quad (45)$$

To finish the proof, we can now supply an upper bound on our one-sample gradient estimation. By the Cauchy-Schwartz inequality, the desired is obtained:

$$\begin{aligned} \left| \widehat{\nabla}_i^t \Phi(x_i) \right| &= \left| \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i,\varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t \right| = |v_i^t| \cdot \left| \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i,\varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) \right| \\ &\leq |v_i^t| \cdot \|\boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \cdot \|\boldsymbol{\psi}_i(\mathbf{x}^t)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{n^2}{\rho_t}. \end{aligned} \quad (46)$$

□

Our second lemma establishes a universal upper bound on the second-order moment of the one-sample gradient approximation given by $\widehat{\nabla}_i^t \Phi(x_i) = \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^{-1} \psi_i(\mathbf{x}^t) v_i^t$.

LEMMA C.2. *At each time t , the following holds for each agent i and candidate coalition $x_i \in [n]$:*

$$\mathbb{E}^t \left[\widehat{\nabla}_i^t \Phi(x_i)^2 \right] = \mathbb{E}^t \left[(\psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^{-1} \psi_i(\mathbf{x}^t) v_i^t)^2 \right] \leq \frac{n^3}{\rho_t}, \quad (47)$$

where $\mathbb{E}^t[\cdot]$ is the expectation over $x_i^t \sim \varphi^t$ for each agent i and all the utilities' randomness at time t .

PROOF. Recalling that $(v_i^t)^2 \leq n^2$, one can observe that:

$$\begin{aligned} \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[(\psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}^t) v_i^t)^2 \right] &\leq n^2 \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[(\psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}^t))^2 \right] \\ &= n^2 \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}^t) \psi_i(\mathbf{x}^t)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}_{-i}^t, x_i) \right] \\ &= n^2 \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[\psi_i(\mathbf{x}^t) \psi_i(\mathbf{x}^t)^\top \right] (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}_{-i}^t, x_i), \end{aligned} \quad (48)$$

where in the last equality we used the fact that only the quantity $\psi_i(\mathbf{x}^t) \psi_i(\mathbf{x}^t)^\top$ depends on $x_i^t \sim \varphi_i^t$. Recalling that the covariance matrix at time t is given by $\Psi_{i, \varphi_i^t}^t = \mathbb{E}_{x_i^t \sim \varphi_i^t} [\psi_i(\mathbf{x}_{-i}^t, x_i) \psi_i(\mathbf{x}_{-i}^t, x_i)^\top]$, we have that:

$$\begin{aligned} \mathbb{E}_{x_i^t \sim \varphi_i^t} \left[(\psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}^t) v_i^t)^2 \right] &\leq n^2 \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \psi_i(\mathbf{x}_{-i}^t, x_i) = \\ &= n^2 \|\psi_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i, \varphi_i^t}^t)^{-1}}^2 \leq \frac{n^3}{\rho_t}, \end{aligned} \quad (49)$$

where the last inequality is by (44). As such, (47) readily follows from (49). $\square$

We are now ready to prove Lemma 4.2. The proof of our lemma leverages the following well-known results concerning Bernstein's inequality:

LEMMA C.3. (Bernstein's Inequality) *Let $\mathcal{R}_1, \dots, \mathcal{R}_K$ be K i.i.d. random variables and set $\bar{\mathcal{R}}_K = \frac{1}{K} \sum_{k \in [K]} \mathcal{R}_k$. For any $k \in [K]$, assume that there is a positive constant $c > 0$ such that $|\mathcal{R}_k| \leq c$ and denote $\sigma^2 := \text{Var}(\mathcal{R}_k)$. Then, for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$:*

$$|\bar{\mathcal{R}}_K - \mathbb{E}[\mathcal{R}_k]| \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{K}} + \frac{2c \log(1/\delta)}{3K}. \quad (50)$$

In particular, if $\sigma^2 \leq \frac{2c^2 \log(1/\delta)}{9n}$, then, with probability at least $1 - \delta$, it holds that:

$$|\bar{\mathcal{R}}_K - \mathbb{E}[\mathcal{R}_k]| \leq \frac{4c \log(1/\delta)}{K}. \quad (51)$$

LEMMA 4.2. *At each time t and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:*

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} \right| \leq O \left(\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (52)$$

In particular, if $\rho_t \geq \frac{2 \log(1/\delta)}{9}$, then, with probability at least $1 - \delta$, it holds that:

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} \right| \leq O \left(\frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (53)$$

PROOF. Recall that our one-sample gradient estimator is given by:

$$\widehat{\nabla}_i^t \Phi(x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t. \quad (54)$$

As such, denoting $\mathcal{R}_t = \widehat{\nabla}_i^t \Phi(x_i) = \boldsymbol{\psi}_i(\mathbf{x}_{-i}^t, x_i)^\top (\Psi_{i, \varphi_i^t}^t)^{-1} \boldsymbol{\psi}_i(\mathbf{x}^t) v_i^t$ for each time $t \in [T]$, then we can apply Lemma C.3 to $\mathcal{R}_t$. Further, by Lemma A.3, $\widehat{\nabla}_i^t \Phi(x_i)$ is an unbiased estimator of the potential's gradient, meaning that $\mathbb{E}^t[\widehat{\nabla}_i^t \Phi(x_i)] = \mathbb{E}[\mathcal{R}_k] = (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i}$. For any time $t \in [T]$, observe that $|\mathcal{R}_t| \leq \frac{n^2}{\rho_t}$ by Lemma C.1. Moreover, $\mathbb{E}^t[\mathcal{R}_t^2] \leq \frac{n^3}{\rho_t}$ by Lemma C.2, and thereby $\text{Var}(\mathcal{R}) = \mathbb{E}^t[\mathcal{R}_t^2] - \mathbb{E}^t[\mathcal{R}_t]^2 \leq \frac{n^3}{\rho_t}$. Accordingly, for each fixed time t and for any $\delta \in (0, 1]$ we can apply Bernstein's inequality as in (50) to obtain that the following holds with probability at least $1 - \delta$ for each fixed agent i and fixed candidate coalition $x_i \in [n]$:

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\boldsymbol{\varphi}^t))_{x_i} \right| \leq \sqrt{\frac{2n^3 \log(1/\delta)}{\rho_t}} + \frac{2n^2 \log(1/\delta)}{3\rho_t}. \quad (55)$$

One can then apply a union bound to obtain (52). Further, note that for using (51) so to attain (53), we should enforce that:

$$\frac{n^3}{\rho_t} \leq \frac{2n^4 \log(1/\delta)}{9\rho_t^2 n} \iff \rho_t \leq \frac{2 \log(1/\delta)}{9}. \quad (56)$$

as required. $\square$

D Proof of Lemma 4.3

LEMMA 4.3. *At each time $t \geq 3$ and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i :*

$$\begin{aligned} \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty &\leq O\left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \right. \\ &\quad \left. + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right). \end{aligned} \quad (57)$$

In particular, if $\rho_t \leq \frac{2 \log(1/\delta)}{9}$ for any time t , then, with probability at least $1 - \delta$, it holds that:

$$\|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty \leq O\left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (58)$$

Further, for $t \in \{1, 2\}$ the error is bounded as in Lemma 4.2 since those time steps satisfy $\|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty = \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi\|_\infty$.

PROOF. First, recall that agent i sets her gradient estimation $\mathbf{g}_i^t$ at time $t \geq 2$ to be a weighted average involving the previous gradient estimate $\mathbf{g}_i^{t-1}$ and the one-sample estimation $\widehat{\nabla}_i^t \Phi$:

$$\mathbf{g}_i^t = (1 - \mu_t) \left(\mathbf{g}_i^{t-1} + \widehat{\Delta}_i^t \right) + \mu_t \widehat{\nabla}_i^t \Phi, \quad (59)$$

where $\mu_t \in (0, 1)$ is an averaging parameter and $\widehat{\Delta}_i^t = \widehat{\nabla}_i^t \Phi - \widehat{\nabla}_i^{t-1} \Phi$ is an unbiased estimator of the gradient variation $\Delta_i^t = \nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})$. Therefore, we denote agent i 's error term corresponding to our gradient estimation in (59) as follows:

$$\mathcal{A}_i^t := \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty, \quad (60)$$

where $\|\cdot\|_\infty$ is the standard L_∞ norm. Note that $\mathcal{A}_i^t$ is equivalent to:

$$\mathcal{A}_i^t = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} + (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})) - (g_i^t - \mathbf{g}_i^{t-1}) \right\|_\infty. \quad (61)$$

Since $g_i^t - \mathbf{g}_i^{t-1} = -\mu_t \mathbf{g}_i^{t-1} + (1 - \mu_t) \widehat{\Delta}_i^t + \mu_t \widehat{\nabla}_i^t \Phi$ by (59) and $\Delta_i^t = \nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})$, then:

$$\mathcal{A}_i^t = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} + \Delta_i^t - (1 - \mu_t) \widehat{\Delta}_i^t - \mu_t \widehat{\nabla}_i^t \Phi + \mu_t \mathbf{g}_i^{t-1} \right\|_\infty \quad (62a)$$

$$= \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} + (1 - \mu_t) (\Delta_i^t - \widehat{\Delta}_i^t) + \mu_t \Delta_i^t - \mu_t \widehat{\nabla}_i^t \Phi + \mu_t \mathbf{g}_i^{t-1} \right\|_\infty \quad (62b)$$

$$= \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} + (1 - \mu_t) [(\nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})) - (\widehat{\nabla}_i^t \Phi - \widehat{\nabla}_i^{t-1} \Phi)] \right. \quad (62c)$$

$$\left. + \mu_t (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi) - \mu_t (\nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1}) \right\|_\infty \quad (62d)$$

$$= \left\| (1 - \mu_t) (\nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1}) + (1 - \mu_t) (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi) \right. \quad (62e)$$

$$\left. + (1 - \mu_t) (\widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})) + \mu_t (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi) \right\|_\infty \quad (62f)$$

$$= \left\| (1 - \mu_t) (\nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1}) + (1 - \mu_t) (\widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})) + (\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi) \right\|_\infty \quad (62g)$$

$$\leq (1 - \mu_t) \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} \right\|_\infty + (1 - \mu_t) \left\| \widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) \right\|_\infty + \left\| \nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi \right\|_\infty \quad (62h)$$

$$= (1 - \mu_t) \mathcal{A}_i^{t-1} + (1 - \mu_t) \left\| \widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) \right\|_\infty + \left\| \nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi \right\|_\infty, \quad (62i)$$

where in (62c) we used the fact that $\widehat{\Delta}_i^t = \widehat{\nabla}_i^t \Phi - \widehat{\nabla}_i^{t-1} \Phi$ and $\Delta_i^t = \nabla_i \Phi(\boldsymbol{\varphi}^t) - \nabla_i \Phi(\boldsymbol{\varphi}^{t-1})$. Further, (62h) is due to the triangle inequality and (62i) stems from our definition in (60), according to which $\mathcal{A}_i^{t-1} = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^{t-1}) - \mathbf{g}_i^{t-1} \right\|_\infty$ holds.

Next, we proceed to proving Lemma 4.3. We provide the proof for (57) while the proof for (58) follows from similar arguments. Initially, we make a few observations regarding the time instants 1 and 2. Recall that for $t = 1$ it holds that $g_i^1 = \widehat{\nabla}_i^1 \Phi$, and thus:

$$\mathcal{A}_i^1 = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^1) - \widehat{\nabla}_i^1 \Phi \right\|_\infty. \quad (63)$$

Thereby, we can use Lemma 4.2 in order to lower bound the right hand side of (63). Specifically, for any $\delta \in (0, 1]$, using (52) from Lemma 4.2 we can obtain that the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $\mathbf{x}_i \in [n]$:

$$\mathcal{A}_i^1 = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^1) - \widehat{\nabla}_i^1 \Phi \right\|_\infty \leq O \left(\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_1}} + \frac{n^3 \log(nT/\delta)}{\rho_1} \right). \quad (64)$$

Similarly, for $t = 2$, note that by (62g) the following is satisfied:

$$\begin{aligned} \mathcal{A}_i^2 &= \left\| (1 - \mu_2) (\nabla_i \Phi(\boldsymbol{\varphi}^1) - \mathbf{g}_i^1) + (1 - \mu_2) (\widehat{\nabla}_i^1 \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^1)) + (\nabla_i \Phi(\boldsymbol{\varphi}^2) - \widehat{\nabla}_i^2 \Phi) \right\|_\infty \\ &= \left\| (1 - \mu_2) (\nabla_i \Phi(\boldsymbol{\varphi}^1) - \widehat{\nabla}_i^1 \Phi) + (1 - \mu_2) (\widehat{\nabla}_i^1 \Phi - \nabla_i \Phi(\boldsymbol{\varphi}^1)) + (\nabla_i \Phi(\boldsymbol{\varphi}^2) - \widehat{\nabla}_i^2 \Phi) \right\|_\infty = \left\| \nabla_i \Phi(\boldsymbol{\varphi}^2) - \widehat{\nabla}_i^2 \Phi \right\|_\infty. \end{aligned} \quad (65)$$

Accordingly, by (52) from Lemma 4.2 and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\mathcal{A}_i^2 = \left\| \nabla_i \Phi(\varphi^2) - \widehat{\nabla}_i^2 \Phi \right\|_\infty \leq O \left(\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_2}} + \frac{n^3 \log(nT/\delta)}{\rho_2} \right). \quad (66)$$

Therefore, using (62) and the above observations, we next prove by induction on $t \geq 3$ that, for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\mathcal{A}_i^t \leq O \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right). \quad (67)$$

For the base case where $t = 3$, by combining (62) with (66) and (52) from Lemma 4.2, we indeed obtain the desired:

$$\mathcal{A}_i^3 \leq O \left((1 - \mu_3) \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_2}} + \frac{n^3 \log(nT/\delta)}{\rho_2} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_3}} + \frac{n^3 \log(nT/\delta)}{\rho_3} \right), \quad (68)$$

as desired. For $t = 4$, we can obtain in a similar manner that:

$$\begin{aligned} \mathcal{A}_i^4 &\leq O \left((1 - \mu_4) \left[(1 - \mu_3) \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_2}} + \frac{n^3 \log(nT/\delta)}{\rho_2} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_3}} + \frac{n^3 \log(nT/\delta)}{\rho_3} \right] \right. \\ &\quad \left. + (1 - \mu_4) \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_3}} + \frac{n^3 \log(nT/\delta)}{\rho_3} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_4}} + \frac{n^3 \log(nT/\delta)}{\rho_4} \right) = \\ &= O \left((1 - \mu_3)(1 - \mu_4) \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_2}} + \frac{n^3 \log(nT/\delta)}{\rho_2} \right] + (1 - \mu_4) \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_3}} + \frac{n^3 \log(nT/\delta)}{\rho_3} \right] \right. \\ &\quad \left. + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_4}} + \frac{n^3 \log(nT/\delta)}{\rho_4} \right) \\ &= O \left(\sum_{\tau_1=2}^3 \left[\prod_{\tau_2=\tau_1+1}^4 (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_4}} + \frac{n^3 \log(nT/\delta)}{\rho_4} \right). \end{aligned} \quad (69)$$

For the inductive step, we assume that (67) holds for $t - 1$ and prove it for t . Hence, similarly to the base case, for any $\delta \in (0, 1]$, we obtain that the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$ due to the inductive assumption:

$$\begin{aligned} \mathcal{A}_i^t &\leq (1 - \mu_t) \mathcal{A}_i^{t-1} + (1 - \mu_t) \left\| \widehat{\nabla}_i^{t-1} \Phi - \nabla_i \Phi(\varphi^{t-1}) \right\|_\infty + \left\| \nabla_i \Phi(\varphi^t) - \widehat{\nabla}_i^t \Phi \right\|_\infty \\ &\leq O \left((1 - \mu_t) \left[\sum_{\tau_1=2}^{t-2} \left[\prod_{\tau_2=\tau_1+1}^{t-1} (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} \right] \right] \right. \\ &\quad \left. + (1 - \mu_t) \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{t-1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{t-1}} + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right) \\ &= O \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^4 \log(nT/\delta)}{\rho_t}} + \frac{n^3 \log(nT/\delta)}{\rho_t} \right), \end{aligned} \quad (70)$$

as required. In order to prove (58), note that we should use (53) instead of (52) from Lemma 4.2 in the above proof. $\square$

E Proof of Theorem 4.1

Before proving Theorem 4.1, we first observe that:

LEMMA E.1. *The potential function $\Phi(\boldsymbol{\varphi}) = \sum_{i \in N} V_i(\boldsymbol{\varphi})$ is n^2 -smooth and n^2 -Lipschitz with respect to the L_1 -norm defined as $\|\mathbf{z}\|_1 = \sum_{i=1}^n z_i$ for any vector $\mathbf{z} \in \mathbb{R}^n$.*

PROOF. Recall that $\Phi(\boldsymbol{\varphi}) = \sum_{i \in N} V_i(\boldsymbol{\varphi})$ and $V_i(\boldsymbol{\varphi}) = \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\varphi}}[d_i(\mathbf{x})]$. Therefore, for any two joint strategies $\boldsymbol{\varphi}, \boldsymbol{\varphi}'$, the following holds by the triangle inequality:

$$\begin{aligned} |\Phi(\boldsymbol{\varphi}) - \Phi(\boldsymbol{\varphi}')| &\leq \sum_{i \in N} |V_i(\boldsymbol{\varphi}) - V_i(\boldsymbol{\varphi}')| = \sum_{i \in N} \left| \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\varphi}}[d_i(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \boldsymbol{\varphi}'}[d_i(\mathbf{x})] \right| \\ &= \sum_{i \in N} \left| \sum_{\mathbf{x} \in [0,1]^n} \boldsymbol{\varphi}(\mathbf{x}) d_i(\mathbf{x}) - \sum_{\mathbf{x} \in [0,1]^n} \boldsymbol{\varphi}'(\mathbf{x}) d_i(\mathbf{x}) \right| = \sum_{i \in N} \left| \sum_{\mathbf{x} \in [0,1]^n} [\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}'(\mathbf{x})] d_i(\mathbf{x}) \right| \quad (71) \\ &\leq \sum_{i \in N} \sum_{\mathbf{x} \in [0,1]^n} |\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}'(\mathbf{x})| |d_i(\mathbf{x})| \leq n^2 \sum_{\mathbf{x} \in [0,1]^n} |\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}'(\mathbf{x})| = n^2 \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1, \end{aligned}$$

where we used the fact that $|d_i(\mathbf{x})| \leq n$ and the definition of the L_1 -norm. This proves that the potential function $\Phi(\boldsymbol{\varphi})$ is indeed n^2 -smooth.

For Lipschitzness, recall that we proved in the full paper that $\nabla_{\varphi_i^t} \Phi(x_i) = \mathbb{E}_{\mathbf{x}_{-i} \sim \boldsymbol{\varphi}_{-i}^t} [d_i(\mathbf{x}_{-i}, x_i)]$ for each agent i at any time t . Thereby, by arguments similar to (71), we can obtain that $\|\nabla_{\boldsymbol{\varphi}} \Phi - \nabla_{\boldsymbol{\varphi}'} \Phi\|_\infty \leq n^2 \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1$, meaning that the potential function $\Phi(\boldsymbol{\varphi})$ is indeed n^2 -Lipschitz. $\square$

We are now ready to prove Theorem 4.1 which provides the Nash regret bound of D1S-FW (Algorithm 2):

THEOREM 4.1. *For any $\delta \in (0, 1]$, D1S-FW (Algorithm 2) with $\rho_t = \frac{1}{T}$, $\eta_t = \frac{1}{\sqrt{T}}$, $\mu_t = 1 - \frac{1}{T}$ for any time t and $T \geq 2$ obtains the following Nash regret bound with probability at least $1 - \delta$:*

$$\mathcal{R}^T \leq O\left((n + n^3 + n^4)\sqrt{T} + n^4 \log(nT/\delta)\right). \quad (72)$$

If $\delta \leq e^{-\frac{9}{2T}}$, then, with probability at least $1 - \delta$, the Nash regret bound in (72) can be reduced to:

$$\mathcal{R}^T \leq O\left((n + n^3 + n^4)\sqrt{T} + n^4 \log(nT/\delta)\right). \quad (73)$$

PROOF. Since the potential function $\Phi(\boldsymbol{\varphi}) = \sum_{i \in N} V_i(\boldsymbol{\varphi})$ is n^2 -smooth with respect to the L_1 -norm $\|\cdot\|_1$ by Lemma E.1, we infer that:

$$\Phi(\boldsymbol{\varphi}^{t+1}) \geq \Phi(\boldsymbol{\varphi}^t) + \langle \nabla \Phi(\boldsymbol{\varphi}^t), \boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t \rangle - \frac{n^2}{2} \|\boldsymbol{\varphi}^{t+1} - \boldsymbol{\varphi}^t\|_1^2. \quad (74)$$

By definition, recall that $\varphi_i^{t+1} = (1 - \rho_t) \left[\varphi_i^t + \eta_t (\hat{\phi}_i^t - \varphi_i^t) \right] + \rho_t \pi_i^t = (1 - \rho_t) \left[(1 - \eta_t) \varphi_i^t + \eta_t \hat{\phi}_i^t \right] + \rho_t \pi_i^t$. Therefore, letting $\hat{\phi}^t = (\hat{\phi}_i^t)_{i \in N}$ and $\pi^t = (\pi_i^t)_{i \in N}$, we obtain from (74) that:

$$\begin{aligned}
\Phi(\varphi^{t+1}) &\geq \Phi(\varphi^t) + (1 - \rho_t)(1 - \eta_t) \langle \nabla \Phi(\varphi^t), \varphi^t - \varphi^t \rangle + (1 - \rho_t) \eta_t \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle \\
&\quad + \rho_t \langle \nabla \Phi(\varphi^t), \pi^t - \varphi^t \rangle - \frac{n^2}{2} \left\| (1 - \rho_t) [\varphi^t + \eta_t (\hat{\phi}^t - \varphi^t)] + \rho_t \pi^t - \varphi^t \right\|_1^2 \\
&= \Phi(\varphi^t) + (1 - \rho_t) \eta_t \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle \\
&\quad - \rho_t \langle \nabla \Phi(\varphi^t), \varphi^t - \pi^t \rangle - \frac{n^2}{2} \left\| (1 - \rho_t) \eta_t (\hat{\phi}^t - \varphi^t) + \rho_t (\pi^t - \varphi^t) \right\|_1^2 \geq \\
&\geq \Phi(\varphi^t) + (1 - \rho_t) \eta_t \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle \\
&\quad - \rho_t \|\nabla \Phi(\varphi^t)\|_\infty \cdot \|\varphi^t - \pi^t\|_1 - \frac{n^2}{2} \left[(1 - \rho_t)^2 \eta_t^2 \|\hat{\phi}^t - \varphi^t\|_1^2 + \rho_t^2 \|\pi^t - \varphi^t\|_1^2 \right],
\end{aligned} \tag{75}$$

where the last inequality follows from $\langle \nabla \Phi(\varphi^t), \pi^t - \varphi^t \rangle \leq \|\nabla \Phi(\varphi^t)\|_\infty \cdot \|\pi^t - \varphi^t\|_1$ by the Cauchy-Schwartz inequality and applying triangle inequality to the term $\|(1 - \rho_t) \eta_t (\hat{\phi}^t - \varphi^t) + \rho_t (\pi^t - \varphi^t)\|_1^2$. Noting that $\|\varphi - \varphi'\|_1 \leq n$ for any two joint strategies $\varphi, \varphi' \in \mathcal{S}_n^n$ and $\|\nabla \Phi(\varphi^t)\|_\infty \leq n^2$ as Φ is n^2 -Lipschitz by Lemma E.1, we infer from (75) that:

$$\Phi(\varphi^{t+1}) \geq \Phi(\varphi^t) + (1 - \rho_t) \eta_t \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle - \rho_t \cdot n^3 - \frac{n^4}{2} \left[(1 - \rho_t)^2 \eta_t^2 + \rho_t^2 \right]. \tag{76}$$

Consequently, we would like to devise a lower bound on the term $\langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle$ in (76). Recall that, for each agent i , $\hat{\phi}_i^t$ at time t is computed according to the classical Frank-Wolfe update via $\hat{\phi}_i^t = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \widehat{\nabla}_i^{m,K} \Phi \rangle$. As such, we denote the *true* strategy optimizing this linear optimization step as $\phi_i^t = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \nabla \Phi(\varphi^t) \rangle$ and let $\phi^t := (\phi_i^t)_{i \in N}$ be the induced joint strategy. Moreover, as discussed in Appendix B, recall that the Frank-Wolfe gap $G(\varphi)$ of the potential function that is associated with a joint strategy φ is given by $G(\varphi) = \max_{\varphi' \in \mathcal{S}_n^n} \langle \varphi' - \varphi, \nabla \varphi \Phi \rangle$. We also denote the gradient estimator of $\nabla \Phi(\varphi^t)$ by $\mathbf{g}^t = (\mathbf{g}_i^t)_{i \in [n]}$. Given those notations, note that:

$$\langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle = \langle \nabla \Phi(\varphi^t) - \mathbf{g}^t + \mathbf{g}^t, \hat{\phi}^t - \varphi^t \rangle = \tag{77a}$$

$$= \langle \mathbf{g}^t, \hat{\phi}^t - \varphi^t \rangle - \langle \mathbf{g}^t - \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle \tag{77b}$$

$$\geq \langle \mathbf{g}^t, \hat{\phi}^t - \varphi^t \rangle - \|\mathbf{g}^t - \nabla \Phi(\varphi^t)\|_\infty \|\hat{\phi}^t - \varphi^t\|_1 \tag{77c}$$

$$= \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle - \langle \nabla \Phi(\varphi^t) - \mathbf{g}^t, \hat{\phi}^t - \varphi^t \rangle - \|\mathbf{g}^t - \nabla \Phi(\varphi^t)\|_\infty \|\hat{\phi}^t - \varphi^t\|_1 \tag{77d}$$

$$\geq \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle - \|\nabla \Phi(\varphi^t) - \mathbf{g}^t\|_\infty \|\hat{\phi}^t - \varphi^t\|_1 - \|\nabla \Phi(\varphi^t) - \mathbf{g}^t\|_\infty \|\hat{\phi}^t - \varphi^t\|_1 \tag{77e}$$

$$\geq \langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle - 2n \|\nabla \Phi(\varphi^t) - \mathbf{g}^t\|_\infty, \tag{77f}$$

where the inequality in (77c) is by the Cauchy-Schwartz inequality and the fact that $\hat{\phi}_i^t = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \widehat{\nabla}_i^{m,K} \Phi \rangle$ for each agent i , meaning that $\langle \mathbf{g}^t, \hat{\phi}^t - \varphi^t \rangle \geq \langle \mathbf{g}^t, \phi^t - \varphi^t \rangle$. Further, (77e) is by the Cauchy-Schwartz inequality, while the inequality in (77f) stems from $\|\hat{\phi}^t - \varphi^t\|_1 \leq n$. Now, since $G(\varphi) = \max_{\varphi' \in \mathcal{S}_n^n} \langle \varphi' - \varphi, \nabla \varphi \Phi \rangle$ and the strategy ϕ_i^t satisfies $\phi_i^t = \arg \max_{\phi \in \mathcal{S}_n} \langle \phi, \nabla \Phi(\varphi^t) \rangle$ for each agent i , then $\langle \nabla \Phi(\varphi^t), \phi^t - \varphi^t \rangle \geq G(\varphi^t)$. Combined with (77), we infer that:

$$\langle \nabla \Phi(\varphi^t), \hat{\phi}^t - \varphi^t \rangle \geq G(\varphi^t) - 2n \|\nabla \Phi(\varphi^t) - \mathbf{g}^t\|_\infty. \tag{78}$$

Accordingly, we can use Lemma 4.3 so as to lower bound the right hand side of (78). Specifically, at each window m and for any $\delta \in (0, 1]$, using (57) from Lemma 4.3 we can obtain that the following holds with probability at

least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\begin{aligned} \langle \nabla \Phi(\boldsymbol{\varphi}^t), \hat{\boldsymbol{\phi}}^t - \boldsymbol{\varphi}^t \rangle &\geq G(\boldsymbol{\varphi}^t) - \mathcal{O} \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] \right. \\ &\quad \left. + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right). \end{aligned} \quad (79)$$

Combining (76) with (79), we have that:

$$\begin{aligned} \Phi(\boldsymbol{\varphi}^{t+1}) &\geq \Phi(\boldsymbol{\varphi}^t) + (1 - \rho_t)\eta_t G(\boldsymbol{\varphi}^t) - \rho_t \cdot n^3 - \frac{n^4}{2} [(1 - \rho_t)^2 \eta_t^2 + \rho_t^2] \\ &\quad - (1 - \rho_t)\eta_t \cdot \mathcal{O} \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right). \end{aligned} \quad (80)$$

From which we infer the following upper bound on the Frank-Wolfe gap:

$$\begin{aligned} G(\boldsymbol{\varphi}^t) &\leq \frac{\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^t)}{(1 - \rho_t)\eta_t} \\ &\quad + \mathcal{O} \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right) \\ &\quad + \frac{\rho_t n^3}{(1 - \rho_t)\eta_t} + \frac{n^4 (1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4}{2(1 - \rho_t)\eta_t}. \end{aligned} \quad (81)$$

By summing over any time $t \in [T]$, we obtain that the Nash regret is upper bounded as follows:

$$\mathcal{R}^T = \sum_{t \in [T]} G(\boldsymbol{\varphi}^t) \leq \frac{\sum_{t \in [T]} [\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^t)]}{(1 - \rho_t)\eta_t} + \frac{\rho_t n^3 T}{(1 - \rho_t)\eta_t} + \frac{n^4 T(1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4 T}{2(1 - \rho_t)\eta_t} \quad (82a)$$

$$+ O\left(\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right] \right) \quad (82b)$$

$$= \frac{\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^1)}{(1 - \rho_t)\eta_t} + \frac{\rho_t n^3 T}{(1 - \rho_t)\eta_t} + \frac{n^4 T(1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4 T}{2(1 - \rho_t)\eta_t} \quad (82c)$$

$$+ O\left(\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right] \right) \quad (82d)$$

$$\leq \frac{n}{(1 - \rho_t)\eta_t} + \frac{\rho_t n^3 T}{(1 - \rho_t)\eta_t} + \frac{n^4 T(1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4 T}{2(1 - \rho_t)\eta_t} \quad (82e)$$

$$+ O\left(\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} \right] + \sqrt{\frac{n^6 \log(nT/\delta)}{\rho_t}} + \frac{n^4 \log(nT/\delta)}{\rho_t} \right] \right), \quad (82f)$$

where for the equality in (82c) we used the fact that $\sum_{t \in [T]} [\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^t)] = \Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^1)$ since this is a telescoping series, and the inequality in (82e) is due to $|\Phi(\boldsymbol{\varphi}^{t+1}) - \Phi(\boldsymbol{\varphi}^1)| \leq n$. As such, given a constant $\mu \in (0, 1]$, we select $\rho_t = \frac{1}{T}$, $\eta_t = \frac{1}{\sqrt{T}}$ and $\mu_t = \mu$ for any time t . Noting that for $T \geq 2$ it holds that $1 \geq 1 - \rho_t \geq \frac{1}{2}$, we infer from (82) that the following holds with probability at least $1 - \delta$:

$$\begin{aligned} \mathcal{R}^T &= \sum_{t \in [T]} G(\boldsymbol{\varphi}^t) \leq \\ &\leq O\left(n \cdot \sqrt{T} + \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\frac{1}{T}}} + \frac{n^4 \log(nT/\delta)}{\frac{1}{T}} \right] \sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu) \right] + 1 \right] \right. \\ &\quad \left. + \frac{\frac{1}{T} \cdot n^3 T}{\frac{1}{\sqrt{T}}} + \frac{n^4 T \cdot \frac{1}{\sqrt{T}}}{2} + \frac{\frac{1}{T^2} n^4 T}{2 \cdot \frac{1}{\sqrt{T}}} \right) \quad (83) \\ &= O\left(n \cdot \sqrt{T} + \left[\sqrt{\frac{n^6 \log(nT/\delta)}{\frac{1}{T}}} + \frac{n^4 \log(nT/\delta)}{\frac{1}{T}} \right] \sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1 - \mu)^{t-\tau_1} + 1 \right] + n^3 \cdot \sqrt{T} + \frac{n^4 \sqrt{T}}{2} + \frac{n^4}{2\sqrt{T}} \right). \end{aligned}$$

Hence, we would like to come up with a closed form for the term $\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1-\mu)^{t-\tau_1} + 1 \right]$. Indeed, note that:

$$\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1-\mu)^{t-\tau_1} + 1 \right] = T + \sum_{t \in [T]} (1-\mu)^t \sum_{\tau_1=2}^{t-1} \frac{1}{(1-\mu)^{\tau_1}} = T + \sum_{t \in [T]} (1-\mu)^t \cdot \frac{\frac{1}{(1-\mu)^2} \left(1 - \frac{1}{(1-\mu)^{t-2}}\right)}{1 - \frac{1}{1-\mu}} \quad (84a)$$

$$= T + \sum_{t \in [T]} (1-\mu)^t \cdot \frac{\frac{1}{(1-\mu)^2} - \frac{1}{(1-\mu)^t}}{\frac{\mu}{1-\mu}} = T + \sum_{t \in [T]} (1-\mu)^t \cdot \frac{1}{\mu} \left(\frac{1}{1-\mu} - \frac{1}{(1-\mu)^{t-1}} \right) \quad (84b)$$

$$= T + \sum_{t \in [T]} \frac{1}{\mu} [(1-\mu)^{t-1} - (1-\mu)^{-1}] = T - \frac{T}{\mu(1-\mu)} + \frac{1}{\mu(1-\mu)} \sum_{t \in [T]} (1-\mu)^t \quad (84c)$$

$$= T - \frac{T}{\mu(1-\mu)} + \frac{1}{\mu(1-\mu)} \cdot \frac{(1-\mu)[1 - (1-\mu)^T]}{1 - (1-\mu)} = T - \frac{T}{\mu(1-\mu)} + \frac{1}{\mu(1-\mu)} \cdot \frac{(1-\mu)[1 - (1-\mu)^T]}{\mu} \quad (84d)$$

$$= T - \frac{T}{\mu(1-\mu)} + \frac{1 - (1-\mu)^T}{\mu^2} \leq T - \frac{T}{\mu(1-\mu)} + \frac{1 - (1-\mu)^2}{\mu^2}, \quad (84e)$$

where in (84a) we used the fact that $\sum_{\tau_1=2}^{t-1} \frac{1}{(1-\mu)^{\tau_1}}$ is a geometric series, (84c) stems from $\sum_{t \in [T]} (1-\mu)^t$ also being a geometric series and the inequality in (84e) is due to $T \geq 2$. By setting $\mu = 1 - \frac{1}{T}$, we obtain that:

$$\begin{aligned} & \sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1-\mu)^{t-\tau_1} + 1 \right] \\ & \leq T - \frac{T}{\mu(1-\mu)} + \frac{1 - (1-\mu)^2}{\mu^2} = T - \frac{T}{\frac{1}{T}(1 - \frac{1}{T})} + \frac{1 - \frac{1}{T^2}}{(1 - \frac{1}{T})^2} = T - \frac{T^2}{\frac{T-1}{T}} + \frac{1 - \frac{1}{T^2}}{\frac{(T-1)^2}{T^2}} \\ & = T - \frac{T^3}{T-1} + \frac{T^2 - 1}{(T-1)^2} = \frac{T(T-1)^2 - T^3(T-1) + T^2 - 1}{(T-1)^2} \\ & = \frac{T(T^2 - 2T + 1) - T^4 + T^3 + T^2 - 1}{(T-1)^2} = \frac{2T^3 - T^2 + T - T^4 - 1}{T^2 - 2T + 1}. \end{aligned} \quad (85)$$

We claim that $\frac{2T^3 - T^2 + T - T^4 - 1}{T^2 - 2T + 1} \leq \frac{1}{T}$. Indeed, this is equivalent to the following:

$$\begin{aligned} & 2T^4 - T^3 + T^2 - T^5 - T \leq T^2 - 2T + 1 \iff \\ & 0 \geq 2T^4 - T^3 - T^5 + T - 1 = T^4 - T^3 - 1 - T^5 + T^4 + T \\ & = T^4 - T^3 - 1 - T(T^4 - T^3 - 1) = (1-T)(T^4 - T^3 - 1) \end{aligned} \quad (86)$$

One can then easily verify that the first root of $(1-T)(T^4 - T^3 - 1)$ are 1 and the other two roots are approximately $-0.81917, 1, 1.3803$. For $T = 1.2$, note that $(1-T)(T^4 - T^3 - 1) = 0.13088 > 0$, which means that $(1-T)(T^4 - T^3 - 1)$ is decreasing as a function of T for any $T \geq 1.3803$ as 1.3803 is its root. In particular, for our choice of η , we obtained that $\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1-\mu)^{t-\tau_1} + 1 \right]$ in (83) satisfies:

$$\sum_{t \in [T]} \left[\sum_{\tau_1=2}^{t-1} (1-\mu)^{t-\tau_1} + 1 \right] \leq \frac{1}{T}. \quad (87)$$

Combining (87) with (83), we conclude that:

$$\begin{aligned}\mathcal{R}^T &= \sum_{t \in [T]} G(\boldsymbol{\varphi}^t) \leq O\left(n \cdot \sqrt{T} + \left\lceil \sqrt{\frac{n^6 \log(nT/\delta)}{\frac{1}{T}}} + \frac{n^4 \log(nT/\delta)}{\frac{1}{T}} \right\rceil \cdot \frac{1}{T} + n^3 \cdot \sqrt{T} + \frac{n^4 \sqrt{T}}{2} + \frac{n^4}{2\sqrt{T}}\right) \\ &= O\left((n + n^3 + n^4) \sqrt{T} + n^4 \log(nT/\delta)\right).\end{aligned}\quad (88)$$

Subsequently, we show that the same Nash regret bound is obtained by applying (53) from Lemma 4.2 to (78) instead of (52) as done earlier, even though (53) provides us with a smaller error. For any $\delta \in (0, 1]$, when choosing $\rho_t = \frac{1}{T}$, recall that for (53) in Lemma 4.2 to hold with probability at least $1 - \delta$ we require that $\rho_t \leq \frac{2 \log(1/\delta)}{9}$, which is equivalent to $\delta \leq e^{-\frac{9}{2T}}$. Namely, noting that $e^{-\frac{9}{2T}} \leq 1$ for $T \geq 2$, we infer that at each time t and for any $\delta \in (0, 1]$ with $\delta \leq e^{-\frac{9}{2T}}$, using (53) from Lemma 4.2 we can obtain that the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\langle \nabla \Phi(\boldsymbol{\varphi}^t), \hat{\boldsymbol{\varphi}}^t - \boldsymbol{\varphi}^t \rangle \geq G(\boldsymbol{\varphi}^t) - O\left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} + \frac{n^4 \log(nT/\delta)}{\rho_t}\right). \quad (89)$$

By arguments similar to the ones that yield (82), we can infer that the Nash regret is upper bounded as:

$$\begin{aligned}\mathcal{R}^T &= \sum_{t \in [T]} G(\boldsymbol{\varphi}^t) \leq \frac{n}{(1 - \rho_t)\eta_t} + O\left(\sum_{t \in [T]} \sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \frac{n^4 \log(nT/\delta)}{\rho_{\tau_1}} + \frac{n^4 \log(nT/\delta)}{\rho_t}\right) \\ &\quad + \frac{\rho_t n^3 T}{(1 - \rho_t)\eta_t} + \frac{n^4 T (1 - \rho_t)\eta_t}{2} + \frac{\rho_t^2 n^4 T}{2(1 - \rho_t)\eta_t}.\end{aligned}\quad (90)$$

Hence, by selecting $\rho_t = \frac{1}{T}$, $\eta_t = \frac{1}{\sqrt{T}}$ and $\mu_t = 1 - \frac{1}{T}$, notice that for $T \geq 2$ it holds that $1 \geq 1 - \rho_t \geq \frac{1}{2}$, and thus we infer from (90) that the following holds with probability at least $1 - \delta$:

$$\begin{aligned}\mathcal{R}^T &= \sum_{t \in [T]} G(\boldsymbol{\varphi}^t) \leq O\left(n \cdot \sqrt{T} + \frac{n^4 \log(nT/\delta)}{\frac{1}{T}} \cdot \frac{1}{T} + \frac{\frac{1}{T} \cdot n^3 T}{\frac{1}{\sqrt{T}}} + \frac{n^4 T \cdot \frac{1}{\sqrt{T}}}{2} + \frac{\frac{1}{T^2} n^4 T}{2 \frac{1}{\sqrt{T}}}\right) \\ &= O\left(n \cdot \sqrt{T} + n^4 \log(nT/\delta) + n^3 \cdot \sqrt{T} + \frac{n^4 \sqrt{T}}{2} + \frac{n^4}{2\sqrt{T}}\right) \\ &= O\left((n + n^3 + n^4) \sqrt{T} + n^4 \log(nT/\delta)\right).\end{aligned}\quad (91)$$

Note that we reduced the dependency on $T^{5/6}$ by a magnitude of n^3 . $\square$

F The Multi-Sample Gradient Estimator is Unbiased

In this section, we show that our multi-sample gradient estimation can be easily proved to be unbiased. Our proof follows arguments similar to the ones used for proving that our one-sample gradient estimator is unbiased (See Lemma A.3).

LEMMA F.1. *At any window $m \in [M]$, the multi-sample estimation $\widehat{\nabla}_i^{m,K} \Phi(x_i)$ of the gradient $(\nabla_i \Phi(\boldsymbol{\varphi}^m))_{x_i}$ is unbiased for each agent i and selection $x_i \in [n]$, i.e.:*

$$\mathbb{E}^m \left[\widehat{\nabla}_i^{m,K} \Phi(x_i) \right] = (\nabla_i \Phi(\boldsymbol{\varphi}^m))_{x_i}. \quad (92)$$

where $\mathbb{E}^m[\cdot]$ is the expectation over $x_i^{m,k} \sim \varphi_i^m$ for each step $k \in [K]$ and each agent i as well as all the utilities' randomness at window m .

PROOF. First, recall that our estimator is given by $\widehat{\nabla}_i^{m,K} \Phi(x_i) = \frac{1}{K} \sum_{k \in [K]} \langle \psi_i(\mathbf{x}_{-i}^{m,k}, x_i), \hat{\xi}_i^{m,k} \rangle$ of $(\nabla_i \Phi(\boldsymbol{\varphi}^m))_{x_i}$, which satisfies the following:

$$\widehat{\nabla}_i^{m,K} \Phi(x_i) = \frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}) v_i^{m,k}. \quad (93)$$

Now, during the k th step of window m , recall that the utility feedback $v_i^{m,k}$ is agent i 's utility from the coalition she joined. As such, if $v_{i,j}^{m,k}$ is agent i 's utility from each agent $j \in N$, then $v_i^{m,k} = \sum_{j \in \pi^{m,k}(i)} v_{i,j}^{m,k}$. By the definition of $\psi_i(\mathbf{x}^{m,k})$, we infer that $v_i^{m,k} = \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}$. Therefore, given that $\mathbf{x}_{-i}^{m,k}$ is the joint selection of all agents other than agent i during any step $k \in [K]$ of window m , then after taking the expectation over $x_i^{m,k'} \sim \varphi_i^m$ for each step $k' \in [K]$ in (26), we have that:

$$\begin{aligned} \mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\widehat{\nabla}_i^{m,K} \Phi(x_i)] &= \mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} \left[\frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}) v_i^{m,k} \right] \\ &= \mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} \left[\frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N} \right]. \end{aligned} \quad (94)$$

Therefore, by the linearity of the expectation, the following holds:

$$\begin{aligned} &\mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\widehat{\nabla}_i^{m,K} \Phi(x_i)] \\ &= \frac{1}{K} \sum_{k \in [K]} \mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}]. \end{aligned} \quad (95)$$

For any $k \in [K]$, since $\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}$ is independent of $x_i^{m,k'}$ for each $k' \neq k$, then:

$$\begin{aligned} &\mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}] \\ &= \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}]. \end{aligned} \quad (96)$$

Further, notice that:

$$\begin{aligned} &\mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top [v_{i,j}^{m,k}]_{j \in N}] \\ &= \mathbb{E}_{x_i^{m,1} \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,1}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,1})^{-1} \psi_i(\mathbf{x}^{m,1}) \psi_i(\mathbf{x}^{m,1})^\top [v_{i,j}^{m,1}]_{j \in N}]. \end{aligned} \quad (97)$$

Combining (97) with (96), we infer that:

$$\mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\widehat{\nabla}_i^{m,K} \Phi(x_i)] = \mathbb{E}_{x_i^{m,1} \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,1}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,1})^{-1} \psi_i(\mathbf{x}^{m,1}) \psi_i(\mathbf{x}^{m,1})^\top [v_{i,j}^{m,1}]_{j \in N}]. \quad (98)$$

Noting that only the quantity $\psi_i(\mathbf{x}^{m,1}) \psi_i(\mathbf{x}^{m,1})^\top$ depends on $x_i^{m,1} \sim \varphi_i^m$, then:

$$\begin{aligned} &\mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\widehat{\nabla}_i^{m,K} \Phi(x_i)] = \\ &= \psi_i(\mathbf{x}_{-i}^{m,1}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,1})^{-1} \mathbb{E}_{x_i^{m,1} \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,1}, x_i^{m,1}) \psi_i(\mathbf{x}_{-i}^{m,1}, x_i^{m,1})^\top]_{j \in N} [v_{i,j}^{m,1}]_{j \in N}. \end{aligned} \quad (99)$$

Recalling that the covariance matrix during the k th step of window m is $\Psi_{i, \varphi_i^m}^{m,k} = \mathbb{E}_{x_i \sim \varphi_i^m} [\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top]$, we have that:

$$\mathbb{E}_{x_i^{m,k'} \sim \varphi_i^m \forall k' \in [K]} [\widehat{\nabla}_i^{m,K} \Phi(x_i)] = \psi_i(\mathbf{x}_{-i}^{m,1}, x_i)^\top [v_{i,j}^{m,1}]_{j \in N}. \quad (100)$$

Now, recall that $\psi_i(\mathbf{x}_{-i}^{m,1}, x_i) \in \{0, 1\}^n$ is given by $[\psi_i(\mathbf{x}_{-i}^{m,1}, x_i)]_j = \mathbb{1}\{x_j^{m,1} = x_i\}$ for any agent $j \in N$, capturing whether agent j joined agent i 's selected coalition or not. This means that $\mathbb{E}_{x_j^{m,1} \sim \varphi_j^m} [\mathbb{1}\{x_j^{m,1} = x_i\}] = \varphi_j^m(x_i)$. Hence, by taking the expectation over $\mathbf{x}_{-i}^{m,k} \sim \boldsymbol{\varphi}_{-i}^m$, we have that:

$$\mathbb{E}_{\mathbf{x}_{-i}^{m,k} \sim \boldsymbol{\varphi}_{-i}^m \forall k' \in [K]} \left[\widehat{\nabla}_i^{m,K} \Phi(x_i) \right] = \boldsymbol{\varphi}^m(x_i)^\top [v_{i,j}^{m,k}]_{j \in N}, \quad (101)$$

where we used our definition of $\boldsymbol{\varphi}^m(x_i) \in [0, 1]^n$ as $[\boldsymbol{\varphi}^m(x_i)]_j = \varphi_j^m(x_i)$ for any agent $j \in N$. Finally, by also taking the expectation over the utilities' randomness, then we conclude that:

$$\mathbb{E}^m \left[\widehat{\nabla}_i^{m,K} \Phi(x_i) \right] = \boldsymbol{\varphi}^m(x_i)^\top \xi_i = (\nabla_i \Phi(\boldsymbol{\varphi}^m))_{x_i}, \quad (102)$$

where we used the fact that $(\nabla_i \Phi(\boldsymbol{\varphi}^m))_{x_i} = \langle \boldsymbol{\varphi}^m(x_i), \xi_i \rangle$ by equation (3) in the full paper. $\square$

G Proof of Lemma 5.1

Before presenting the proof of Lemma 5.1, we require two auxiliary results which are similar to Lemmas C.1 and C.2 from Appendix C. The first lemma illustrates a universal upper bound on the gradient approximation that is induced by the selected strategy at each time instant.

LEMMA G.1. *During any step k of each window m , the following holds for each agent i and candidate coalition $x_i \in [n]$:*

$$\left| \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}) v_i^{m,k} \right| \leq \frac{n^2}{\rho}, \quad (103)$$

and in particular:

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) \right| = \left| \frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}) v_i^{m,k} \right| \leq \frac{n^2}{\rho}. \quad (104)$$

PROOF. Recall that agent i 's strategy φ_i^m at window m is given as regular Frank-Wolfe update with the exploration strategy π_i^m , i.e., $\varphi_i^m = (1 - \rho)[\varphi_i^{m-1} + \eta(\phi_i^{m-1} - \varphi_i^{m-1})] + \rho\pi_i^{m-1}$. The covariance matrix $\Psi_{i,\varphi_i^m}^{m,k} = \mathbb{E}_{x_i \sim \varphi_i^m} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top \right]$ at the k th step of window m therefore satisfies the following:

$$\begin{aligned} \Psi_{i,\varphi_i^m}^{m,k} &= \mathbb{E}_{x_i \sim \varphi_i^m} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top \right] \\ &= (1 - \rho) \mathbb{E}_{x_i \sim [\varphi_i^{m-1} + \eta(\phi_i^{m-1} - \varphi_i^{m-1})]} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top \right] + \rho \mathbb{E}_{x_i \sim \pi_i^{m-1}} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top \right] \\ &= (1 - \rho) \Psi_{i,\tilde{\varphi}_i^{m-1}}^{m,k} + \rho \Psi_{i,\pi_i^{m-1}}^{m,k}, \end{aligned} \quad (105)$$

where we denoted $\tilde{\varphi}_i^{m-1} := \varphi_i^{m-1} + \eta(\phi_i^{m-1} - \varphi_i^{m-1})$.

Further, recall that $\pi_i^m = \frac{1}{K} \sum_{k \in [K]} \pi_i^{m,k}$, where $\pi_i^{m,k}$ is agent i 's G-optimal design for $\{\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)\}_{x_i \in [n]}$ during step k of window m , which satisfies $\pi_i^{m,k} \in \arg \min_{\phi \in \Delta([n])} g(\phi)$ for:

$$g(\phi) = \max_{x_i \in [n]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\phi}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i). \quad (106)$$

Combining $\pi_i^{m-1} = \frac{1}{K} \sum_{k \in [K]} \pi_i^{m-1,k}$ with (105), we infer that:

$$\Psi_{i,\varphi_i^m}^{m,k} = (1 - \rho) \Psi_{i,\tilde{\varphi}_i^{m-1}}^{m,k} + \rho \cdot \frac{1}{K} \sum_{k \in [K]} \Psi_{i,\pi_i^{m-1,k}}^{m,k}. \quad (107)$$

Accordingly, for each $\mathbf{z} \in \mathbb{R}^n$:

$$\mathbf{z}^\top \Psi_{i,\varphi_i^m}^{m,k} \mathbf{z} = (1 - \rho) \cdot \mathbf{z}^\top \Psi_{i,\tilde{\varphi}_i^{m-1}}^{m,k} \mathbf{z} + \rho \cdot \frac{1}{K} \sum_{k \in [K]} \mathbf{z}^\top \Psi_{i,\pi_i^{m-1,k}}^{m,k} \mathbf{z} \geq \rho \cdot \frac{1}{K} \sum_{k \in [K]} \mathbf{z}^\top \Psi_{i,\pi_i^{m-1,k}}^{m,k} \mathbf{z} \geq \rho \cdot \mathbf{z}^\top \Psi_{i,\pi_i^{m-1,\tilde{k}}}^{m,k} \mathbf{z}, \quad (108)$$

where $\tilde{k} \in \arg \min_{k \in [K]} \mathbf{z}^\top \Psi_{i,\pi_i^{m-1,k}}^{m,k} \mathbf{z}$. We thus infer that:

$$\psi_i(\mathbf{x}^{m,k})^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \leq \frac{1}{\rho} \psi_i(\mathbf{x}^{m,k})^\top (\Psi_{i,\pi_i^{m-1,\tilde{k}}}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}). \quad (109)$$

By the Kiefer–Wolfowitz Theorem (Lattimore and Szepesvári 2020, Chapter 21), it is known that:

$$g(\pi_i^{m-1,k}) = \max_{x_i \in [n]} \psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i)^\top (\Psi_{i,\pi_i^{m-1,k}}^{m-1,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i) = n. \quad (110)$$

In particular, note that:

$$\begin{aligned} n &= \max_{\mathbf{x}_{-i}^{m-1,k} \in [n]^{n-1}} g(\pi_i^{m-1,k}) \\ &= \max_{\mathbf{x}_{-i}^{m-1,k} \in [n]^{n-1}} \max_{x_i \in [n]} \psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i)^\top (\mathbb{E}_{x_i \sim \pi_i^{m-1}} [\psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i)^\top])^{-1} \psi_i(\mathbf{x}_{-i}^{m-1,k}, x_i) \\ &\geq \psi_i(\mathbf{x}^{m,k})^\top (\Psi_{i,\pi_i^{m-1,\tilde{k}}}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}). \end{aligned} \quad (111)$$

Applying (111) to (109) yields:

$$\psi_i(\mathbf{x}^{m,k})^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \leq \frac{n}{\rho}. \quad (112)$$

Note that this expression is essentially the norm of $\psi_i(\mathbf{x}^{m,k})$ with respect to the norm induced by the matrix $(\Psi_{i,\varphi_i^m}^{m,k})^{-1}$. Specifically, for any matrix $A \in \mathbb{R}^{n \times n}$ and two vectors $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$, the inner product $\langle \cdot, \cdot \rangle_A$ induced by A is given as $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle_A := \mathbf{z}_1^\top A \mathbf{z}_2$, which reduces to the norm $\| \cdot \|_A$ induced by A via $\|\mathbf{z}_1\|_A^2 = \langle \mathbf{z}_1, \mathbf{z}_1 \rangle_A$. That is, (112) indicates that:

$$\|\psi_i(\mathbf{x}^{m,k})\|_{(\Psi_{i,\varphi_i^m}^{m,k})^{-1}}^2 \leq \frac{n}{\rho} \iff \|\psi_i(\mathbf{x}^{m,k})\|_{(\Psi_{i,\varphi_i^m}^{m,k})^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho}}. \quad (113)$$

By similar arguments, note that for any $x_i \in [n]$:

$$\|\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)\|_{(\Psi_{i,\varphi_i^m}^{m,k})^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho}}. \quad (114)$$

To finish the proof, we can now supply an upper bound on each summand in (104). By the Cauchy-Schwartz inequality, observe that:

$$\begin{aligned} \left| \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k} \right| &= |v_i^{m,k}| \cdot \left| \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i,\varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \right| \\ &\leq |v_i^{m,k}| \cdot \|\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)\|_{(\Psi_{i,\varphi_i^m}^{m,k})^{-1}} \cdot \|\psi_i(\mathbf{x}^{m,k})\|_{(\Psi_{i,\varphi_i^m}^{m,k})^{-1}} \leq \frac{n^2}{\rho}, \end{aligned} \quad (115)$$

where in the last inequality we used (113) and (114) as well as the fact that $|v_i^{m,k}| \leq n$. By the triangle inequality, we conclude that:

$$\begin{aligned} \left| \widehat{\nabla}_i^{m,K} \Phi(x_i) \right| &= \left| \frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k} \right| \\ &\leq \frac{1}{K} \sum_{k \in [K]} |\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k}| \leq \frac{n^2}{\rho}, \end{aligned} \quad (116)$$

as desired. $\square$

Our second lemma illustrates a universal upper bound on the second-order moment of each summand in the multi-sample gradient approximation.

LEMMA G.2. *During any step $k \in [K]$ of window $m \in [M]$, the following holds for each agent i and candidate coalition $x_i \in [n]$:*

$$\mathbb{E}^m \left[(\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k})^2 \right] \leq \frac{n^3}{\rho}, \quad (117)$$

where $\mathbb{E}^m[\cdot]$ is the expectation over $x_i^{m,k} \sim \varphi^m$ for each step $k \in [K]$ and each agent i as well as all the utilities' randomness at window m .

PROOF. Recalling that $(v_i^{m,k})^2 \leq n^2$, one can observe that:

$$\begin{aligned} \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} \left[(\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k})^2 \right] &\leq n^2 \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} \left[(\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}))^2 \right] \\ &= n^2 \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \right] \\ &= n^2 \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} \left[\psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top \right] (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i), \end{aligned} \quad (118)$$

where in the last equality we used the fact that only the quantity $\psi_i(\mathbf{x}^{m,k}) \psi_i(\mathbf{x}^{m,k})^\top$ depends on $x_i^{m,k} \sim \varphi_i^m$. Recalling that the covariance matrix during the k th step of window m is given by $\Psi_{i, \varphi_i^m}^{m,k} = \mathbb{E}_{x_i \sim \varphi_i^m} \left[\psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top \right]$, we have that:

$$\begin{aligned} \mathbb{E}_{x_i^{m,k} \sim \varphi_i^m} \left[(\psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k})^2 \right] &\leq n^2 \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i) \\ &= n^2 \|\psi_i(\mathbf{x}^{m,k})\|_{(\Psi_{i, \varphi_i^m}^{m,k})^{-1}}^2 \leq \frac{n^3}{\rho}, \end{aligned} \quad (119)$$

where the last inequality is by (113). As such, (117) readily follows from (119). $\square$

We are now ready to prove Lemma 5.1 from the full paper:

LEMMA 5.1. *At each window m and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:*

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) - (\nabla_i \Phi(\varphi^m))_{x_i} \right| \leq O \left(\sqrt{\frac{n^4 \log(nM/\delta)}{\rho K}} + \frac{n^3 \log(nM/\delta)}{\rho K} \right). \quad (120)$$

In particular, if $\rho \leq \frac{2 \log(1/\delta)}{9}$, then, with probability at least $1 - \delta$, it holds that:

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) - (\nabla_i \Phi(\varphi^m))_{x_i} \right| \leq O \left(\frac{n^3 \log(nM/\delta)}{\rho K} \right). \quad (121)$$

PROOF. Recall that our multi-sample gradient estimator is given by:

$$\widehat{\nabla}_i^{m,K} \Phi(x_i) = \frac{1}{K} \sum_{k \in [K]} \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k}. \quad (122)$$

As such, denoting $\mathcal{R}_k = \psi_i(\mathbf{x}_{-i}^{m,k}, x_i)^\top (\Psi_{i, \varphi_i^m}^{m,k})^{-1} \psi_i(\mathbf{x}^{m,k}) v_i^{m,k}$ for each step $k \in [K]$, then $\widehat{\nabla}_i^{m,K} \Phi(x_i) = \bar{\mathcal{R}}_K$ under the notations of Lemma C.3. Further, by Lemma F.1, $\widehat{\nabla}_i^{m,K} \Phi(x_i)$ is an unbiased estimator of the potential's gradient, meaning that $\mathbb{E}^m \left[\widehat{\nabla}_i^{m,K} \Phi(x_i) \right] = \mathbb{E}[\mathcal{R}_k] = (\nabla_i \Phi(\varphi^m))_{x_i}$. For any step $k \in [K]$, observe that $|\mathcal{R}_k| \leq \frac{n^2}{\rho}$ by Lemma G.1. Moreover, $\mathbb{E}^m[\mathcal{R}_k^2] \leq \frac{n^3}{\rho}$ by Lemma G.2, and thereby $\text{Var}(\mathcal{R}) = \mathbb{E}^m[\mathcal{R}_k^2] - \mathbb{E}^m[\mathcal{R}_k]^2 \leq \frac{n^3}{\rho}$. Accordingly, for each fixed window m and for any $\delta \in (0, 1]$ we can apply Bernstein's inequality as in (50) to obtain that the following holds with probability at least $1 - \delta$ for each fixed agent i and fixed candidate coalition $x_i \in [n]$:

$$\left| \widehat{\nabla}_i^{m,K} \Phi(x_i) - (\nabla_i \Phi(\varphi^m))_{x_i} \right| \leq \sqrt{\frac{2n^3 \log(1/\delta)}{\rho K}} + \frac{2n^2 \log(1/\delta)}{3\rho K}. \quad (123)$$

One can then apply a union bound to obtain (120). Further, note that for using (51) so to attain (121), we should enforce that:

$$\frac{n^3}{\rho} \leq \frac{2n^4 \log(1/\delta)}{9\rho^2 n} \iff \rho \leq \frac{2 \log(1/\delta)}{9}, \quad (124)$$

as required. $\square$

H Proof of Theorem 5.1

THEOREM 5.1. Consider $T = MK$ rounds and let $\delta \in (0, 1]$. We distinguish two regimes.

- **General regime:** DMS-FW (Algorithm 3) with $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $M \geq 2$ and $K = M^2$ (so that $T = M^3$) obtains the following Nash regret bound with probability at least $1 - \delta$:

$$\mathcal{R}^T \leq O \left(\left(n + 2n^3 + n^4 + \sqrt{n^6 \mathcal{F}_\delta^T} \right) T^{5/6} + n^4 T^{2/3} \left(1 + \mathcal{F}_\delta^T \right) \right), \quad (125)$$

where $\mathcal{F}_\delta^T := \log(nT^{1/3}/\delta)$. Moreover, the choice $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $K = M^2$ is asymptotically optimal for the general bound: no choice of the window/step split $T = MK$ and step sizes $\rho, \eta \in (0, 1)$ makes the upper bound proved for DMS-FW smaller than $\tilde{O}(T^{5/6})$ (see Lemma I.2 in Appendix I).

- **Reduced regime:** If $\delta \leq e^{-\frac{9}{2M}}$, then the split $K = M^2$ is no longer optimal; the optimal choice is $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$, $M \geq 2$ and $K = M^{3/2}$ (so that $T = M^{5/2}$, i.e., $M = T^{2/5}$), under which, with probability at least $1 - \delta$, the Nash regret bound improves to:

$$\mathcal{R}^T \leq O \left((n + n^3 + n^4) T^{4/5} + n^4 \cdot T^{4/5} \mathcal{G}_\delta^T \right), \quad (126)$$

where $\mathcal{G}_\delta^T := \log(nT^{2/5}/\delta)$. This choice is asymptotically optimal for the reduced bound: no choice of $T = MK$ and $\rho, \eta \in (0, 1)$ makes the (reduced) upper bound proved for DMS-FW smaller than $\tilde{O}(T^{4/5})$ (see Lemma I.3 in Appendix I).

PROOF. Since the potential function $\Phi(\varphi) = \sum_{i \in N} V_i(\varphi)$ is n^2 -smooth with respect to the L_1 -norm $\|\cdot\|_1$ by Lemma E.1, we infer that:

$$\Phi(\varphi^{m+1}) \geq \Phi(\varphi^m) + \langle \nabla \Phi(\varphi^m), \varphi^{m+1} - \varphi^m \rangle - \frac{n^2}{2} \|\varphi^{m+1} - \varphi^m\|_1^2. \quad (127)$$

By definition, recall that $\varphi_i^{m+1} = (1-\rho)[\varphi_i^m + \eta(\hat{\phi}_i^m - \varphi_i^m)] + \rho\pi_i^m = (1-\rho)[(1-\eta)\varphi_i^m + \eta\hat{\phi}_i^m] + \rho\pi_i^m$. Therefore, letting $\hat{\phi}^m = (\hat{\phi}_i^m)_{i \in N}$ and $\pi^m = (\pi_i^m)_{i \in N}$, we obtain from (127) that:

$$\begin{aligned}
\Phi(\varphi^{m+1}) &\geq \Phi(\varphi^m) + (1-\rho)(1-\eta)\langle \nabla\Phi(\varphi^m), \varphi^m - \varphi^m \rangle + (1-\rho)\eta\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle \\
&\quad + \rho\langle \nabla\Phi(\varphi^m), \pi^m - \varphi^m \rangle - \frac{n^2}{2}\|(1-\rho)[\varphi^m + \eta(\hat{\phi}^m - \varphi^m)] + \rho\pi^m - \varphi^m\|_1^2 \\
&= \Phi(\varphi^m) + (1-\rho)\eta\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle - \rho\langle \nabla\Phi(\varphi^m), \varphi^m - \pi^m \rangle \\
&\quad - \frac{n^2}{2}\|(1-\rho)\eta(\hat{\phi}^m - \varphi^m) + \rho(\pi^m - \varphi^m)\|_1^2 \geq \\
&\geq \Phi(\varphi^m) + (1-\rho)\eta\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle - \rho\|\nabla\Phi(\varphi^m)\|_\infty \cdot \|\varphi^m - \pi^m\|_1 \\
&\quad - \frac{n^2}{2}\left[(1-\rho)^2\eta^2\|\hat{\phi}^m - \varphi^m\|_1^2 + \rho^2\|\pi^m - \varphi^m\|_1^2\right],
\end{aligned} \tag{128}$$

where the last inequality follows from $\langle \nabla\Phi(\varphi^m), \pi^m - \varphi^m \rangle \leq \|\nabla\Phi(\varphi^m)\|_\infty \cdot \|\pi^m - \varphi^m\|_1$ by the Cauchy-Schwartz inequality and applying triangle inequality to the term $\|(1-\rho)\eta(\hat{\phi}^m - \varphi^m) + \rho(\pi^m - \varphi^m)\|_1^2$. Noting that $\|\varphi - \varphi'\|_1 \leq n$ for any two joint strategies $\varphi, \varphi' \in \Delta([n])^n$ and $\|\nabla\Phi(\varphi^m)\|_\infty \leq n^2$ as Φ is n^2 -Lipschitz by Lemma E.1, we infer from (128) that:

$$\Phi(\varphi^{m+1}) \geq \Phi(\varphi^m) + (1-\rho)\eta\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle - \rho \cdot n^3 - \frac{n^4}{2}[(1-\rho)^2\eta^2 + \rho^2]. \tag{129}$$

Consequently, we would like to devise a lower bound on the term $\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle$ in (129). Recall that, for each agent i , $\hat{\phi}_i^m$ at window m is computed according to the classical Frank-Wolfe update via $\hat{\phi}_i^m = \arg \max_{\phi \in \Delta([n])} \langle \phi, \widehat{\nabla}^{m,K}\Phi \rangle$. As such, we denote the *true* strategy optimizing this linear optimization step as $\phi_i^m = \arg \max_{\phi \in \Delta([n])} \langle \phi, \nabla\Phi(\varphi^m) \rangle$ and let $\phi^m := (\phi_i^m)_{i \in N}$ be the induced joint strategy. Moreover, as discussed in Appendix B, recall that the Frank-Wolfe gap $G(\varphi)$ of the potential function that is associated with a joint strategy φ is given by $G(\varphi) = \max_{\varphi' \in \Delta([n])^n} \langle \varphi' - \varphi, \nabla\Phi(\varphi) \rangle$. Given those notations, note that the following holds:

$$\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle = \langle \nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi + \widehat{\nabla}^{m,K}\Phi, \hat{\phi}^m - \varphi^m \rangle \tag{130a}$$

$$= \langle \widehat{\nabla}^{m,K}\Phi, \hat{\phi}^m - \varphi^m \rangle - \langle \widehat{\nabla}^{m,K}\Phi - \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle \tag{130b}$$

$$\geq \langle \widehat{\nabla}^{m,K}\Phi, \phi^m - \varphi^m \rangle - \|\widehat{\nabla}^{m,K}\Phi - \nabla\Phi(\varphi^m)\|_\infty \|\hat{\phi}^m - \varphi^m\|_1 \tag{130c}$$

$$= \langle \nabla\Phi(\varphi^m), \phi^m - \varphi^m \rangle - \langle \nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi, \phi^m - \varphi^m \rangle - \|\widehat{\nabla}^{m,K}\Phi - \nabla\Phi(\varphi^m)\|_\infty \|\hat{\phi}^m - \varphi^m\|_1 \tag{130d}$$

$$\geq \langle \nabla\Phi(\varphi^m), \phi^m - \varphi^m \rangle - \|\nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi\|_\infty \|\phi^m - \varphi^m\|_1 - \|\nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi\|_\infty \|\hat{\phi}^m - \varphi^m\|_1 \tag{130e}$$

$$\geq \langle \nabla\Phi(\varphi^m), \phi^m - \varphi^m \rangle - 2n\|\nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi\|_\infty, \tag{130f}$$

where the inequality in (130c) is by the Cauchy-Schwartz inequality and $\hat{\phi}_i^m = \arg \max_{\phi \in \Delta([n])} \langle \phi, \widehat{\nabla}^{m,K}\Phi \rangle$ for each agent i , meaning that $\langle \widehat{\nabla}^{m,K}\Phi, \hat{\phi}^m - \varphi^m \rangle \geq \langle \widehat{\nabla}^{m,K}\Phi, \phi^m - \varphi^m \rangle$. Further, (130e) is also by the Cauchy-Schwartz inequality, while the inequality in (130f) stems from $\|\hat{\phi}^m - \varphi^m\|_1 \leq n$. Now, since $G(\varphi) = \max_{\varphi' \in \Delta([n])^n} \langle \varphi' - \varphi, \nabla\Phi(\varphi) \rangle$ and the strategy ϕ^m satisfies $\phi_i^m = \arg \max_{\phi \in \Delta([n])} \langle \phi, \nabla\Phi(\varphi^m) \rangle$ for each agent i , then $\langle \nabla\Phi(\varphi^m), \phi^m - \varphi^m \rangle \geq G(\varphi^m)$. Combined with (130), we infer that:

$$\langle \nabla\Phi(\varphi^m), \hat{\phi}^m - \varphi^m \rangle \geq G(\varphi^m) - 2n\|\nabla\Phi(\varphi^m) - \widehat{\nabla}^{m,K}\Phi\|_\infty. \tag{131}$$

Accordingly, we can use Lemma 5.1 so as to lower bound the right hand side of (131). Specifically, at each window m and for any $\delta \in (0, 1]$, using (120) from Lemma 5.1 we can obtain that the following holds with probability at

least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\langle \nabla \Phi(\boldsymbol{\varphi}^m), \hat{\boldsymbol{\phi}}^m - \boldsymbol{\varphi}^m \rangle \geq G(\boldsymbol{\varphi}^m) - O\left(\sqrt{\frac{n^6 \log(nM/\delta)}{\rho K}} + \frac{n^4 \log(nM/\delta)}{\rho K}\right). \quad (132)$$

Combining (129) with (132), we have that:

$$\begin{aligned} \Phi(\boldsymbol{\varphi}^{m+1}) &\geq \Phi(\boldsymbol{\varphi}^m) + (1 - \rho)\eta G(\boldsymbol{\varphi}^m) - (1 - \rho)\eta \cdot O\left(\sqrt{\frac{n^6 \log(nM/\delta)}{\rho K}} + \frac{n^4 \log(nM/\delta)}{\rho K}\right) \\ &\quad - \rho \cdot n^3 - \frac{n^4}{2} [(1 - \rho)^2 \eta^2 + \rho^2]. \end{aligned} \quad (133)$$

From which we infer the following upper bound on the Frank-Wolfe gap:

$$\begin{aligned} G(\boldsymbol{\varphi}^m) &\leq \frac{\Phi(\boldsymbol{\varphi}^{m+1}) - \Phi(\boldsymbol{\varphi}^m)}{(1 - \rho)\eta} + O\left(\sqrt{\frac{n^6 \log(nM/\delta)}{\rho K}} + \frac{n^4 \log(nM/\delta)}{\rho K}\right) \\ &\quad + \frac{\rho n^3}{(1 - \rho)\eta} + \frac{n^4(1 - \rho)\eta}{2} + \frac{\rho^2 n^4}{2(1 - \rho)\eta}. \end{aligned} \quad (134)$$

By summing over any window $m \in [M]$:

$$\sum_{m \in [M]} G(\boldsymbol{\varphi}^m) \leq \frac{\sum_{m \in [M]} [\Phi(\boldsymbol{\varphi}^{m+1}) - \Phi(\boldsymbol{\varphi}^m)]}{(1 - \rho)\eta} + O\left(\sqrt{\frac{n^6 M^2 \log(nM/\delta)}{\rho K}} + \frac{n^4 M \log(nM/\delta)}{\rho K}\right) \quad (135a)$$

$$+ \frac{\rho n^3 M}{(1 - \rho)\eta} + \frac{n^4 M(1 - \rho)\eta}{2} + \frac{\rho^2 n^4 M}{2(1 - \rho)\eta} \quad (135b)$$

$$= \frac{\Phi(\boldsymbol{\varphi}^{M+1}) - \Phi(\boldsymbol{\varphi}^1)}{(1 - \rho)\eta} + O\left(\sqrt{\frac{n^6 M^2 \log(nM/\delta)}{\rho K}} + \frac{n^4 M \log(nM/\delta)}{\rho K}\right) \quad (135c)$$

$$+ \frac{\rho n^3 M}{(1 - \rho)\eta} + \frac{n^4 M(1 - \rho)\eta}{2} + \frac{\rho^2 n^4 M}{2(1 - \rho)\eta} \quad (135d)$$

$$\leq \frac{n}{(1 - \rho)\eta} + O\left(\sqrt{\frac{n^6 M^2 \log(nM/\delta)}{\rho K}} + \frac{n^4 M \log(nM/\delta)}{\rho K}\right) \quad (135e)$$

$$+ \frac{\rho n^3 M}{(1 - \rho)\eta} + \frac{n^4 M(1 - \rho)\eta}{2} + \frac{\rho^2 n^4 M}{2(1 - \rho)\eta}, \quad (135f)$$

where for the equality in (135d) we used the fact that $\sum_{m \in [M]} [\Phi(\boldsymbol{\varphi}^{m+1}) - \Phi(\boldsymbol{\varphi}^m)] = \Phi(\boldsymbol{\varphi}^{M+1}) - \Phi(\boldsymbol{\varphi}^1)$ since this is a telescoping series, and the inequality in (135f) is due to $|\Phi(\boldsymbol{\varphi}^{m+1}) - \Phi(\boldsymbol{\varphi}^1)| \leq n$. As such, by selecting $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$ and $K = M^2$, observe that for $M \geq 2$ it holds that $1 \geq 1 - \rho \geq \frac{1}{2}$, and thus we infer from (135) that the

following holds with probability at least $1 - \delta$:

$$\begin{aligned}
\sum_{m \in [M]} G(\boldsymbol{\varphi}^m) &\leq O \left(n \cdot \sqrt{M} + \sqrt{\frac{n^6 M^2 \log(nM/\delta)}{\frac{1}{M} \cdot M^2}} + \frac{n^4 M \log(nM/\delta)}{\frac{1}{M} \cdot M^2} + \frac{\frac{1}{M} \cdot n^3 M}{\frac{1}{\sqrt{M}}} + \frac{n^4 M \cdot \frac{1}{\sqrt{M}}}{2} + \frac{\frac{1}{M^2} n^4 M}{2 \frac{1}{\sqrt{M}}} \right) \\
&= O \left(n \cdot \sqrt{M} + \sqrt{M} \cdot \sqrt{n^6 \log(nM/\delta)} + n^4 \log(nM/\delta) + n^3 \cdot \sqrt{M} + \frac{n^4 \sqrt{M}}{2} + \frac{n^4}{2\sqrt{M}} \right) \\
&= O \left(\left(n + 2n^3 + n^4 + \sqrt{n^6 \log(nM/\delta)} \right) \sqrt{M} + n^4 (1 + \log(nM/\delta)) \right).
\end{aligned} \tag{136}$$

Now, recall that the Nash regret is upper bounded by the sum of the Frank-Wolfe gaps across all $T = MK$ rounds by (35), as $K = M^2$ we have that:

$$\begin{aligned}
\mathcal{R}^T &\leq K \cdot \sum_{m \in [M]} G(\boldsymbol{\varphi}^m) \leq O \left(\left(n + 2n^3 + n^4 + \sqrt{n^6 \log(nM/\delta)} \right) M^{5/2} + n^4 \cdot M^2 (1 + \log(nM/\delta)) \right) \\
&= O \left(\left(n + 2n^3 + n^4 + \sqrt{n^6 \log(nT^{1/3}/\delta)} \right) T^{5/6} + n^4 \cdot T^{2/3} (1 + \log(nT^{1/3}/\delta)) \right).
\end{aligned} \tag{137}$$

Subsequently, we discuss how we can significantly reduce the obtained Nash regret bound by applying (121) from Lemma 5.1 to (131) instead of (120) as done earlier. For any $\delta \in (0, 1]$, when choosing $\rho = \frac{1}{M}$, recall that for (121) in Lemma 5.1 to hold with probability at least $1 - \delta$ we require that $\rho \leq \frac{2 \log(1/\delta)}{9}$, which is equivalent to $\delta \leq e^{-\frac{9}{2M}}$. Namely, noting that $e^{-\frac{9}{2M}} \leq 1$ for $M \geq 2$, we infer that at each window m and for any $\delta \in (0, 1]$ with $\delta \leq e^{-\frac{9}{2M}}$, using (121) from Lemma 5.1 we can obtain that the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:

$$\langle \nabla \Phi(\boldsymbol{\varphi}^m), \hat{\boldsymbol{\phi}}^m - \boldsymbol{\varphi}^m \rangle \geq G(\boldsymbol{\varphi}^m) - O \left(\frac{n^4 \log(nM/\delta)}{\rho K} \right). \tag{138}$$

By arguments similar to the ones that yield (135), we can infer that:

$$\sum_{m \in [M]} G(\boldsymbol{\varphi}^m) \leq \frac{n}{(1-\rho)\eta} + O \left(\frac{n^4 M \log(nM/\delta)}{\rho K} \right) + \frac{\rho n^3 M}{(1-\rho)\eta} + \frac{n^4 M (1-\rho)\eta}{2} + \frac{\rho^2 n^4 M}{2(1-\rho)\eta}. \tag{139}$$

Hence, in this reduced regime we select the *optimal* split $\rho = \frac{1}{M}$, $\eta = \frac{1}{\sqrt{M}}$ and $K = M^{3/2}$ (established to be optimal in Lemma I.3 of Appendix I). Notice that for $M \geq 2$ it holds that $1 \geq 1 - \rho \geq \frac{1}{2}$, and thus we infer from (139) that the following holds with probability at least $1 - \delta$:

$$\begin{aligned}
\sum_{m \in [M]} G(\boldsymbol{\varphi}^m) &\leq O \left(n \cdot \sqrt{M} + \frac{n^4 M \log(nM/\delta)}{\frac{1}{M} \cdot M^{3/2}} + \frac{\frac{1}{M} \cdot n^3 M}{\frac{1}{\sqrt{M}}} + \frac{n^4 M \cdot \frac{1}{\sqrt{M}}}{2} + \frac{\frac{1}{M^2} n^4 M}{2 \frac{1}{\sqrt{M}}} \right) \\
&= O \left(n \cdot \sqrt{M} + n^4 \sqrt{M} \log(nM/\delta) + n^3 \cdot \sqrt{M} + \frac{n^4 \sqrt{M}}{2} + \frac{n^4}{2\sqrt{M}} \right) \\
&= O \left(\left(n + n^3 + n^4 \right) \sqrt{M} + n^4 \sqrt{M} \log(nM/\delta) \right).
\end{aligned} \tag{140}$$

Now, recall that the Nash regret is upper bounded by the sum of the Frank-Wolfe gaps across all $T = MK$ rounds by (35); as $K = M^{3/2}$ (so that $T = M^{5/2}$, i.e., $M = T^{2/5}$) we have that:

$$\begin{aligned}\mathcal{R}^T &\leq K \cdot \sum_{m \in [M]} G(\boldsymbol{\varphi}^m) \leq O((n + n^3 + n^4)M^2 + n^4 \cdot M^2 \log(nM/\delta)) \\ &= O\left((n + n^3 + n^4)T^{4/5} + n^4 \cdot T^{4/5} \log(nT^{2/5}/\delta)\right).\end{aligned}\quad (141)$$

Note that this reduced-regime bound of order $T^{4/5}$ strictly improves upon the general-regime bound of order $T^{5/6}$ obtained in (137), by using the optimal split $K = M^{3/2}$ in place of $K = M^2$ (see Appendix I). $\square$

I Optimality of the Parameter Choices for DMS-FW

In this appendix, we prove that the parameter choices used in Theorem 5.1 are *asymptotically optimal* for the corresponding regret bounds. We first record the *parameter-free* upper bound underlying the proof of Theorem 5.1, and then minimize it over all parameters.

A parameter-free bound. As shown in the proof of Theorem 5.1 (Appendix H), combining the smoothness inequality (129) with the Frank-Wolfe-gap lower bound (131), rearranging, summing over $m \in [M]$ (a telescoping series with $|\Phi(\boldsymbol{\varphi}^{M+1}) - \Phi(\boldsymbol{\varphi}^1)| \leq n$), and using $\mathcal{R}^T \leq K \sum_{m \in [M]} G(\boldsymbol{\varphi}^m)$ (within a window all K steps share the strategy $\boldsymbol{\varphi}^m$, each incurring the gap $G(\boldsymbol{\varphi}^m)$), yields, for any $\rho, \eta \in (0, 1)$ and $T = MK$, the parameter-free bound:

$$\mathcal{R}^T \leq \frac{nK}{(1-\rho)\eta} + 2nK \sum_{m \in [M]} \left\| \nabla \Phi(\boldsymbol{\varphi}^m) - \widehat{\nabla}^{m,K} \Phi \right\|_\infty + \frac{\rho n^3 T}{(1-\rho)\eta} + \frac{n^4(1-\rho)\eta T}{2} + \frac{\rho^2 n^4 T}{2(1-\rho)\eta}. \quad (142)$$

Write $\mathcal{F} := \log(nM/\delta)$. Assume $\rho \leq \frac{1}{2}$, so that $\frac{1}{2} \leq 1-\rho \leq 1$. Substituting the general error bound (20) of Lemma 5.1 into (142) and using $\sum_{m \in [M]} (\cdot) \leq M \cdot \max_m (\cdot)$ gives, up to absolute constants,

$$\mathcal{R}^T \leq O\left(\frac{nK}{\eta} + n^3 \sqrt{\mathcal{F}} \frac{M\sqrt{K}}{\sqrt{\rho}} + \frac{n^4 \mathcal{F} T}{\rho K} + \frac{\rho n^3 T}{\eta} + n^4 \eta T + \frac{\rho^2 n^4 T}{\eta}\right), \quad (143)$$

where we used $T = MK$. In the *reduced* regime $\rho \leq \frac{2}{9} \log(1/\delta)$, the improved error bound (21) removes the second (square-root) term, yielding

$$\mathcal{R}^T \leq O\left(\frac{nK}{\eta} + \frac{n^4 \mathcal{F} T}{\rho K} + \frac{\rho n^3 T}{\eta} + n^4 \eta T + \frac{\rho^2 n^4 T}{\eta}\right). \quad (144)$$

Exponent form. We now minimize (143)–(144) over all admissible parameters.

LEMMA I.1. Fix n and δ . Any admissible parameters may be written as $M = T^a$ (so $K = T^{1-a}$), $\rho = T^{-b}$ and $\eta = T^{-c}$ with $a \in (0, 1)$ and $b, c \geq 0$. This parametrization is exact, as for any admissible (K, ρ, η) one can define $a := \log_T M$, $b := -\log_T \rho$ and $c := -\log_T \eta$. Since $\mathcal{F} = \log(nM/\delta) = \tilde{\Theta}(1)$ for $M = \text{poly}(T)$, the bound in (143) equals $\tilde{\Theta}(T^{V_{\text{gen}}})$ and (144) equals $\tilde{\Theta}(T^{V_{\text{red}}})$, where

$$V_{\text{gen}} = \max\{\alpha_1, \alpha_6, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}, \quad V_{\text{red}} = \max\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}, \quad (145)$$

with $\alpha_1 = 1 - a + c$, $\alpha_6 = \frac{1}{2} + \frac{a}{2} + \frac{b}{2}$, $\alpha_2 = a + b$, $\alpha_3 = 1 - b + c$, $\alpha_4 = 1 - c$ and $\alpha_5 = 1 - 2b + c$. Moreover $\alpha_5 \leq \alpha_3$ for all $b \geq 0$.

PROOF. Each summand of (143)–(144) is a monomial in T :

- the first term is $n T^{(1-a)+c}$,
- the square-root term is $n^3 \sqrt{\mathcal{F}} T^{a+(1-a)/2+b/2} = n^3 \sqrt{\mathcal{F}} T^{1/2+a/2+b/2}$,
- the third is $n^4 \mathcal{F} T^{1+b-(1-a)} = n^4 \mathcal{F} T^{a+b}$ (as $T/(\rho K) = T^{a+b}$)

- the fourth is $n^3 T^{1-b+c}$; the fifth is $n^4 T^{1-c}$; and the sixth is $n^4 T^{1-2b+c}$.

Taking the largest exponent gives (145). Finally $\alpha_5 = \alpha_3 - b \leq \alpha_3$ since $b \geq 0$. $\square$

LEMMA I.2 (OPTIMALITY OF $K = M^2$ FOR THE GENERAL BOUND). *It holds that $\min_{a \in (0,1), b, e \geq 0} V_{\text{gen}} = \frac{5}{6}$, which is attained at $(a, b, e) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{6})$, i.e., $M = T^{1/3}$, $K = T^{2/3} = M^2$, $\rho = 1/M$ and $\eta = 1/\sqrt{M}$. Hence, the bound in (143) is minimized at $\tilde{O}(T^{5/6})$, and substituting these values into (143) yields (22).*

PROOF. Since $\alpha_5 \leq \alpha_3$, we have $V_{\text{gen}} = \max\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_6\}$.

We begin by proving achievability. At $(a, b, e) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{6})$: $\alpha_1 = 1 - \frac{1}{3} + \frac{1}{6} = \frac{5}{6}$, $\alpha_3 = 1 - \frac{1}{3} + \frac{1}{6} = \frac{5}{6}$, $\alpha_4 = 1 - \frac{1}{6} = \frac{5}{6}$, $\alpha_6 = \frac{1}{2} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}$ and $\alpha_2 = \frac{2}{3} \leq \frac{5}{6}$. Thus $V_{\text{gen}} = \frac{5}{6}$.

We proceed to proving a matching lower bound. Suppose, for contradiction, that $V_{\text{gen}} < \frac{5}{6}$. Then, $\alpha_4 = 1 - c < \frac{5}{6}$ forces $c > \frac{1}{6}$; $\alpha_1 = 1 - a + c < \frac{5}{6}$ forces $a > \frac{1}{6} + c > \frac{1}{3}$; $\alpha_3 = 1 - b + c < \frac{5}{6}$ forces $b > \frac{1}{6} + c > \frac{1}{3}$. Hence, we infer that $\alpha_6 = \frac{1}{2} + \frac{a+b}{2} > \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$, contradicting $\alpha_6 < \frac{5}{6}$. Therefore, $V_{\text{gen}} \geq \frac{5}{6}$, so the minimum is $\frac{5}{6}$. The four constraints $\alpha_1 = \alpha_3 = \alpha_4 = \alpha_6 = \frac{5}{6}$ have the unique solution above. The substitution producing (22) is then verified term-by-term (see (136) for details): with $M = T^{1/3}$, $K = T^{2/3}$, $\rho = T^{-1/3}$, $\eta = T^{-1/6}$ and $\mathcal{F} = \log(nT^{1/3}/\delta) = \mathcal{F}_\delta^T$, the six terms of (143) equal $nT^{5/6}$, $\sqrt{n^6 \mathcal{F}_\delta^T} T^{5/6}$, $n^4 \mathcal{F}_\delta^T T^{2/3}$, $n^3 T^{5/6}$, $n^4 T^{5/6}$ and $n^4 T^{1/2}$, respectively. $\square$

LEMMA I.3 (OPTIMALITY OF $K = M^{3/2}$ FOR THE REDUCED BOUND). *It holds that $\min_{a \in (0,1), b, e \geq 0} V_{\text{red}} = \frac{4}{5}$, attained at $(a, b, e) = (\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$, i.e., $M = T^{2/5}$, $K = T^{3/5} = M^{3/2}$, $\rho = 1/M$ and $\eta = 1/\sqrt{M}$. Hence (144) is minimized at $\tilde{O}(T^{4/5})$, and substituting these values into (144) yields (23). In particular, the choice $K = M^2$ (i.e., $a = \frac{1}{3}$) gives $V_{\text{red}} = \max\{\frac{5}{6}, \frac{2}{3}\} = \frac{5}{6} > \frac{4}{5}$ and is therefore strictly suboptimal for the reduced bound.*

PROOF. Since $\alpha_5 \leq \alpha_3$, $V_{\text{red}} = \max\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$.

We begin by proving achievability. At $(a, b, e) = (\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$: $\alpha_1 = 1 - \frac{2}{5} + \frac{1}{5} = \frac{4}{5}$, $\alpha_2 = \frac{2}{5} + \frac{2}{5} = \frac{4}{5}$, $\alpha_3 = 1 - \frac{2}{5} + \frac{1}{5} = \frac{4}{5}$ and $\alpha_4 = 1 - \frac{1}{5} = \frac{4}{5}$. Thus $V_{\text{red}} = \frac{4}{5}$.

We proceed to proving a matching lower bound. Suppose, for contradiction, that $V_{\text{red}} < \frac{4}{5}$. Then, $\alpha_4 = 1 - c < \frac{4}{5}$ forces $c > \frac{1}{5}$; $\alpha_1 = 1 - a + c < \frac{4}{5}$ forces $a > \frac{1}{5} + c > \frac{2}{5}$; $\alpha_3 = 1 - b + c < \frac{4}{5}$ forces $b > \frac{1}{5} + c > \frac{2}{5}$. Hence, $\alpha_2 = a + b > \frac{4}{5}$, contradicting $\alpha_2 < \frac{4}{5}$. Therefore, we infer that $V_{\text{red}} \geq \frac{4}{5}$, so the minimum is $\frac{4}{5}$. The four constraints $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{4}{5}$ have the unique solution above. For the substitution: with $M = T^{2/5}$, $K = T^{3/5}$, $\rho = T^{-2/5}$, $\eta = T^{-1/5}$ and $\mathcal{F} = \log(nT^{2/5}/\delta) = \mathcal{G}_\delta^T$, the five terms of (144) equal $nT^{4/5}$, $n^4 \mathcal{G}_\delta^T T^{4/5}$, $n^3 T^{4/5}$, $n^4 T^{4/5}$ and $n^4 T^{2/5}$, respectively, giving $\mathcal{R}^T \leq O((n + n^3 + n^4)T^{4/5} + n^4 T^{4/5} \mathcal{G}_\delta^T)$, which is (23). $\square$

PROPOSITION I.1. *Setting $K = 1$ (hence $M = T$) in DMS-FW, with $\rho = 1/M = 1/T$ and $\eta = 1/\sqrt{M} = 1/\sqrt{T}$, the proven bound (144) is $\Omega(n^4 T^2)$, i.e., superlinear and hence vacuous as a Nash regret bound. More generally, by Lemmas I.2 and I.3, the proven bound is at least $\tilde{\Omega}(T^{4/5})$ for every admissible choice of (M, K, ρ, η) , so DMS-FW's guarantee never matches the $\tilde{O}(\sqrt{T})$ bound of D1S-FW (Theorem 4.1).*

PROOF. With $K = 1$, $M = T$ and $\rho = 1/T$, the third term of (144) equals $\frac{n^4 \mathcal{F} T}{\rho K} = n^4 \mathcal{F} \frac{T}{(1/T) \cdot 1} = n^4 \mathcal{F} T^2$. As all terms of (144) are nonnegative, the proven bound is $\Omega(n^4 T^2)$. This reflects that at $K = 1$ the in-window average (18) contains a single summand and provides no variance reduction, while DMS-FW carries no temporal momentum term (contrast (3) in D1S-FW); the mechanism producing the $\tilde{O}(\sqrt{T})$ bound of D1S-FW is therefore absent. $\square$

Together, Lemmas I.2 and I.3 establish the asymptotic optimality claims of Theorem 5.1: the choice $K = M^2$ minimizes the general bound (143) at $\tilde{O}(T^{5/6})$, while in the reduced regime the choice $K = M^{3/2}$ minimizes (144) at $\tilde{O}(T^{4/5})$, strictly improving upon $K = M^2$.

J Discussion: Using an Approximate G-Optimal Design

For ease of reading in the main body of our paper, we assumed that a G-optimal design can be computed accurately and efficiently. In this section, we show that for our work it is sufficient to compute a c -approximate optimal design with minimal impact on performance. Initially, we consider c -approximate G-optimal design in general (Section J.1), and then discuss near-optimal G-optimal designs (Section J.2).

J.1 Using c -approximate G-optimal designs

We first depict that it suffices for each agent i to use an c -approximate G-optimal design. Beforehand, we provide the rationale behind the minor effects on performance that results from using an approximate G-optimal design. Specifically, we motivate this by initially supplying a brief reminder of our main steps for obtaining our results when using an *exact* G-optimal design. We extensively discuss D1S-FW (Algorithm 2), while the arguments for DMS-FW (Algorithm 3) are similar. For D1S-FW, recall that we derived the upper bounds on our estimators' errors using the Kiefer–Wolfowitz Theorem (Lattimore and Szepesvári 2020, Chapter 21), from which we obtained that agent i 's G-optimal design π_i^t at time t satisfies $g(\pi_i^t) = n$ (as also stated in (41)). In turn, this allowed us to obtain the universal upper bound on the one-sample gradient approximation $\widehat{\nabla}_i^t \Phi(x_i) = \psi_i(\mathbf{x}_{-i}^t, x_i)^\top (\hat{\Psi}_i^t)^{-1} \psi_i(\mathbf{x}^t) v_i^t$ as well as its second-order moment in Lemmas C.1 and C.2. Specifically, we used this property of the G-optimal design to show that $\|\psi_i(\mathbf{x}^t)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho_t}}$ in (44) and $\|\psi_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{\sqrt{n}}{\sqrt{\rho_t}}$ in (45). Then, we proved the correctness of our claims in Lemmas C.1 and C.2, which state that $|\widehat{\nabla}_i^t \Phi(x_i)| \leq |v_i^t| \cdot \|\psi_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \cdot \|\psi_i(\mathbf{x}^t)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}} \leq \frac{n^2}{\rho_t}$ and $\mathbb{E}^t \left[\widehat{\nabla}_i^t \Phi(x_i)^2 \right] \leq n^2 \|\psi_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}}^2 \leq \frac{n^3}{\rho_t}$ (respectively).

Now, we infer the upper bounds on the estimation error that is incurred when each agent i uses an c -approximate G-optimal design $\hat{\pi}_i^t$ at any time t with $f_i^t(\hat{\pi}_i^t) \leq (1+c)n$. Todd (Todd 2016) illustrated that such a design can be computed within $4n(\log \log n + \frac{7}{2}) + \frac{28n}{c}$ iterations by the Frank–Wolfe algorithm with a specific initialization. Accordingly, if we only compute c -approximate G-optimal designs in D1S-FW (Algorithm 2), then the upper bound on the estimation error will now also depend on a factor of $1+c$. Namely, by the above reasoning concerning D1S-FW, the upper bounds in Lemmas C.1 and C.2 will become $|\widehat{\nabla}_i^t \Phi(x_i)| \leq \frac{n^2(1+c)}{\rho_t}$ and $\mathbb{E}^t \left[\widehat{\nabla}_i^t \Phi(x_i)^2 \right] \leq n^2 \|\psi_i(\mathbf{x}_{-i}^t, x_i)\|_{(\Psi_{i,\varphi_i^t}^t)^{-1}}^2 \leq \frac{n^3(1+c)}{\rho_t}$ (respectively). Therefore, by the proof of Lemmas 4.2 and 4.3, we infer the following upper bounds on the estimation error and its second-order moment that are incurred by D1S-FW when only an approximate G-optimal design is employed by each agent:

LEMMA 2 (APPROXIMATE G-OPTIMAL DESIGN). *Assume that each agent i uses an c -approximate G-optimal design $\hat{\pi}_i^t$ at any time t . Then, at each time t and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$ for each agent i and any candidate coalition $x_i \in [n]$:*

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\varphi^t))_{x_i} \right| \leq O \left(\sqrt{\frac{n^4(1+c) \log(nT/\delta)}{\rho_t}} + \frac{n^3(1+c) \log(nT/\delta)}{\rho_t} \right). \quad (146)$$

In particular, if $\rho_t \leq \frac{2 \log(1/\delta)}{9}$, then, with probability at least $1 - \delta$, it holds that:

$$\left| \widehat{\nabla}_i^t \Phi(x_i) - (\nabla_i \Phi(\varphi^t))_{x_i} \right| \leq O \left(\frac{n^3(1+c) \log(nM/\delta)}{\rho_t} \right). \quad (147)$$

LEMMA 3 (APPROXIMATE G-OPTIMAL DESIGN). *Assume that each agent i uses an c -approximate G-optimal design $\hat{\pi}_i^t$ at any time t . Then, at each time $t \geq 3$ and for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$*

for each agent i :

$$\begin{aligned} \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty \leq & \mathcal{O} \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \left[\sqrt{\frac{n^4(1+c)\log(nT/\delta)}{\rho_{\tau_1}}} + \frac{n^3(1+c)\log(nT/\delta)}{\rho_{\tau_1}} \right] \right. \\ & \left. + \sqrt{\frac{n^4(1+c)\log(nT/\delta)}{\rho_t}} + \frac{n^3(1+c)\log(nT/\delta)}{\rho_t} \right). \end{aligned} \quad (148)$$

In particular, if $\rho_t \leq \frac{2\log(1/\delta)}{9}$ for any time t , then, with probability at least $1 - \delta$, it holds that:

$$\|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty \leq \mathcal{O} \left(\sum_{\tau_1=2}^{t-1} \left[\prod_{\tau_2=\tau_1+1}^t (1 - \mu_{\tau_2}) \right] \cdot \frac{n^3(1+c)\log(nT/\delta)}{\rho_{\tau_1}} + \frac{n^3(1+c)\log(nT/\delta)}{\rho_t} \right). \quad (149)$$

Further, for $t \in \{1, 2\}$ the error is bounded as in Lemma 4.2 since those time steps satisfy $\|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \mathbf{g}_i^t\|_\infty = \|\nabla_i \Phi(\boldsymbol{\varphi}^t) - \widehat{\nabla}_i^t \Phi\|_\infty$.

By the proof of Theorems 4.1, when the model's parameters are fixed, recall that the Nash regret bound of D1S-FW (Algorithm 2) is determined by the error incurred by the utilized gradient estimator. As such, we obtain that the Nash regret bound of D1S-FW also depends on a factor of $1 + c$:

THEOREM 1 (APPROXIMATE G-OPTIMAL DESIGN). Assume that each agent i uses an c -approximate G -optimal design $\hat{\pi}_i^t$ at any time t . Then, for any $\delta \in (0, 1]$, D1S-FW (Algorithm 2) with $\rho_t = \frac{1}{T}$, $\eta_t = \frac{1}{\sqrt{T}}$, $\mu_t = 1 - \frac{1}{T}$ for any time t and $T \geq 2$ obtains the following Nash regret bound with probability at least $1 - \delta$:

$$\mathcal{R}^T \leq \tilde{\mathcal{O}} \left((1+c)(n + n^3 + n^4) \sqrt{T} \right). \quad (150)$$

By similar arguments, we can obtain that the Nash regret bound of DMS-FW (Algorithm 3) further depends on a factor of $1 + c$.

J.2 Using Near-optimal G-optimal Designs

A near-optimal G -optimal design was recently introduced (Lattimore, Szepesvari, et al. 2020; Todd 2016). Specifically, Todd (Todd 2016, Chapter 3) showed that, using the Frank-Wolfe algorithm with a proper initialization, each agent i can efficiently compute an approximate optimal design $\hat{\pi}_i^t$ at any time t such that $f_i^t(\hat{\pi}_i^t) \leq 2n$. In fact, Todd illustrated that this can be done within $4n(\log \log n + 21/2)$ iterations of the Frank-Wolfe algorithm using an appropriate initialization. By arguments analogous to Section J.1, we can thus obtain results similar to Lemmas 2 (Approximate G -Optimal Design) and 3 (Approximate G -Optimal Design) as well as Theorem 1 (Approximate G -Optimal Design) by substituting $c = 1$.

J.3 Analysis of D1S-FW's Time Complexity

The algorithm starts with initializing an arbitrary strategy for each agent in $\mathcal{O}(n^2)$ time and then runs over T rounds (line 1). As mentioned in Section 4.3.1, under the assumptions of Theorem 4.1, our algorithm's sample complexity is optimal up to logarithmic factors, that is, the number of rounds required to achieve an ε -Nash stable solution is $T = \mathcal{O}(n^8/\varepsilon^2)$, where $\varepsilon \geq 0$ is the desired accuracy. Each round consists of the following steps:

- (1) Each agent i joins a coalition following her mixed strategy and receives a utility feedback in $\mathcal{O}(n)$ time (lines 3-4), and then she proceeds to the following steps.
- (2) In line 6, agent i then computes a one-sample gradient estimator via equation (8). This is done using linear regression, which takes $\mathcal{O}(n^3)$ time for each agent i due to standard time complexities of matrix inversion and multiplication.

- (3) If $t = 1$, agent i sets the gradient estimation in $O(1)$ time (line 8). Otherwise, if $t \geq 2$, then she computes a one-sample gradient variation estimator in $O(n)$ time (line 10). Afterwards, she computes the gradient estimator $\mathbf{g}_i^t$ using equation 3 in $O(n)$ time (line 11).
- (4) In line 12, agent i optimizes a linear optimization problem to get the feasible ascent strategy as the regular Frank-Wolfe method. This linear optimization step is over the simplex, so its maximum is attained at a vertex of the simplex, i.e., at one of the standard basis vectors. This is equivalent to finding the index of the largest entry in the gradient estimator, which takes $O(n)$ time for each agent i .
- (5) In line 13, agent i then calculates a G-optimal design. As specified in Remark 5, for our purposes it suffices to compute an approximate G-optimal design with minimal impact on performance. Specifically, agent i can *efficiently* compute a near-optimal G-optimal design in $4n(\log \log n + 21/2)$ time (See Appendix J.2 for details).

Each agent requires approximately $O(n^3)$ time per round, resulting in a total time complexity of $O(n^4)$ at each round. Thus, together with the initialization in line 1 and the fact that the number of rounds required to achieve an ε -Nash stable solution is $T = O(n^8/\varepsilon^2)$, our algorithm's time complexity is $O(n^2 + n^{12}/\varepsilon^2)$.

J.4 Analysis of DMS-FW's Time Complexity

The algorithm starts with initializing an arbitrary strategy for each agent in $O(n^2)$ time and then runs over T rounds (line 1). As mentioned in Section 5.3.1, under the (reduced-regime) assumptions of Theorem 5.1, our algorithm's optimal sample complexity is $\tilde{O}(n^{20}/\varepsilon^5)$, i.e., the number of rounds required to achieve an ε -Nash stable solution is $T = \tilde{O}(n^{20}/\varepsilon^5)$, where $\varepsilon \geq 0$ is the desired accuracy. Recall that the number of rounds T is split into $M > 0$ windows, each with $K > 0$ steps. By Theorem 5.1, in this regime we pick the optimal split $K = M^{3/2}$ to obtain the aforementioned sample complexity bound, resulting in $T = M^{5/2}$. As such, the number of windows required to achieve an ε -Nash stable solution is $M = O(n^8/\varepsilon^2)$, while each window consists of $K = M^{3/2} = O(n^{12}/\varepsilon^3)$ steps. Each window consists of the following steps:

- (1) Over K steps, each agent i joins a coalition following her mixed strategy and receives a utility feedback in $O(n)$ time (lines 4-5), and then she proceeds to the following steps. This yields a time complexity of $O(n^{13}/\varepsilon^3)$ for each agent i .
- (2) In line 7, agent i then computes a multi-sample gradient estimator via equation (18). This is done using linear regression for each of the collected K samples, which takes $O(Kn^3) = O(n^{15}/\varepsilon^3)$ time for each agent i due to standard time complexities of matrix (pseudo)inversion and multiplication.
- (3) In line 8, agent i optimizes a linear optimization problem to get the feasible ascent strategy as the regular Frank-Wolfe method. This linear optimization step is over the simplex, so its maximum is attained at a vertex of the simplex, i.e., at one of the standard basis vectors. This is equivalent to finding the index of the largest entry in the gradient estimator, which takes $O(n)$ time for each agent i .
- (4) In line 13, agent i then calculates K G-optimal designs. As specified in Remark 5, for our purposes it suffices to compute an approximate G-optimal design with minimal impact on performance. Specifically, agent i can *efficiently* compute a near-optimal G-optimal design in $4n(\log \log n + 21/2)$ time (See Appendix J.2 for details). Namely, agent i requires $K \cdot 4n(\log \log n + 21/2) = O(4n^{13}(\log \log n + 21/2)/\varepsilon^3)$ time to calculate the K G-optimal designs.

Each agent requires approximately $\tilde{O}(n^{15}/\varepsilon^3)$ time per window, resulting in a total time complexity of $\tilde{O}(n^{16}/\varepsilon^3)$ at each window. Thus, together with the initialization in line 1 and the fact that the number of windows required to achieve an ε -Nash stable solution is $M = O(n^8/\varepsilon^2)$, our algorithm's time complexity is $\tilde{O}(n^2 + n^{24}/\varepsilon^5)$.

Received 07 July 2025; accepted 08 September 2026