DECENTRALIZED ONLINE LEARNING WITH BANDIT FEEDBACK IN FAIR VALUE GAMES

Submission ID: 36085

ABSTRACT

We study decentralized online learning in task-based Fair Value Games (FVGs), where heterogeneous agents repeatedly select tasks and receive their marginal contributions to unknown, *identity-dependent* coalition values. We introduce `OSD-FW`, a projection-free Frank–Wolfe method that uses one sample per update and adapts a known template for *non-convex* objectives. Its new component is a *sampled drift correction*. The centralized form of that correction needs the joint change in every agent's policy, which no individual agent holds, and earlier decentralized methods substitute a compact summary statistic that identity-dependent values do not admit. Ours is built instead from single actions that agents draw and announce for themselves, at a per-round query cost linear in the numbers of agents and tasks. Because the Frank–Wolfe gap equals total exploitability in these games, the resulting stationarity bounds are Nash-regret bounds, with $\mathcal{O}(T^{3/4})$ under unilateral-oracle feedback, $\mathcal{O}(T^{2/3})$ once the correction is added, and $\mathcal{O}(T^{5/6})$ under pure bandit feedback, the last slower than the best rate known for this class. Under a matched value-oracle budget, the drift-corrected variant ranks fourth on average Nash gap but attains the smallest final-iterate gap among the decentralized methods tested.

1 INTRODUCTION

Many multi-agent systems operate without a central coordinator. Examples include robots coordinating a search mission Cai et al. (2025), workers selecting projects in peer-to-peer labor markets Xia et al. (2019), devices sharing bandwidth in ad-hoc wireless networks Zhang et al. (2022), and researchers forming cross-institutional collaborations Marcolino et al. (2013). In these settings, an assignment's value may depend on the members of the coalition and on the task itself. A group may perform one task well and another poorly because skills, complementarities, and interference vary across agents and tasks. Anonymous models, which retain only aggregate quantities such as coalition size or resource load, cannot represent these effects.

Coalition productivity is typically *unknown in advance*. Agents must learn from experience while adapting to the choices of other self-interested agents. Without a global coordinator, each agent acts on local observations and personal feedback. This model avoids a central point of control and applies when agents cannot exchange their policies or private estimates. Fair value games Augustine et al. (2015) and valid utility games Vetta (2002) assign each agent her marginal contribution to a shared welfare function. They are potential games Monderer & Shapley (1996), so they admit pure Nash equilibria and decentralized best-response dynamics. Existing learning results usually assume known coalition values or exploit the structure of congestion games Cui et al. (2022) or additively separable hedonic games Cohen & Agmon (2025b;a). Those assumptions exclude general identity-dependent values.

In this paper, we study decentralized online learning in task-based FVGs with heterogeneous agents and unknown coalition values. Each agent repeatedly selects a task from a personal feasible set and receives her marginal contribution to the realized coalition value. Because values depend jointly on the task and the coalition's membership, the model strictly generalizes congestion games Rosenthal (1973) and additively separable hedonic games Bogomolnaia & Jackson (2002).

Our main contribution is a drift-corrected one-sample estimator that agents can implement without knowing one another's policies when coalition values are identity-dependent. We also show that

the resulting stationarity bound is a Nash-regret bound for this class. Previous decentralized Frank–Wolfe methods construct one-sample gradient or drift estimators from low-dimensional summaries: resource loads in congestion games Cui et al. (2022) and pairwise coefficients in hedonic games Cohen & Agmon (2025a). A general submodular set function of coalition identity has no comparable summary. The centralized construction of Zhang et al. (2020) also assumes access to the joint iterate increment. The sampled correction in Section 3.3.2 avoids both requirements, with variance and per-round cost polynomial in n and m (Proposition 2, Table 7).

The resulting algorithm, `OSD-FW` (One-Sample Decentralized Frank–Wolfe), is projection-free and uses one stochastic sample per update. It requires neither a coordinator nor the exchange of policies or private estimates. Gradient alignment extends from the finite potential game to its multilinear extension, so each Frank–Wolfe direction acts as a decentralized best-response surrogate. Variant A uses a biased momentum estimate and obtains $\mathcal{O}(T^{3/4})$ Nash regret under unilateral-oracle feedback. Forced exploration changes the rate to $\mathcal{O}(T^{5/6})$ under pure bandit feedback. Variant B uses an unbiased drift correction and obtains $\mathcal{O}(T^{2/3})$ Nash regret. Each rate is sublinear and therefore yields convergence to approximate Nash equilibria.

The drift correction of Zhang et al. (2020) multiplies the iterate increment $x^t - x^{t-1}$, which is available to a centralized optimizer. In a game, this increment is the joint policy change, whereas agent i holds only one block. The agent therefore cannot sample from an interpolation of the joint policy. Earlier decentralized Frank–Wolfe methods use compact summaries of the joint policy, namely load counts (Cui et al., 2022) or pairwise coefficients (Cohen & Agmon, 2025a). General identity-dependent set functions do not admit either summary. Section 3.3.2 instead estimates the Jacobian–vector product from two public action draws per agent per round. The Frank–Wolfe update expresses the increment as $\pi_k^t - \pi_k^{t-1} = \eta(\mu_k^{t-1} - \pi_k^{t-1})$, so one sample can estimate the required expectation. The protocol uses $2\lceil \log_2 m \rceil$ bits per agent per round and $O(mn)$ oracle calls, compared with $O(m^2 n)$ for the unsampled form. The public sequence b_k^t, the step size η, and the prescribed initialization determine π_k^t exactly. Thus, the transcript reveals every policy; the result concerns communication and query cost, not privacy (Remark 5).

For arbitrary identity-dependent values, both per-round computation and estimator variance are polynomial in n and m (Proposition 2). The hypothesis class is exponential, so a *model-based* method that tabulates or regresses on v_j can incur exponential cost (Appendix B.5). Table 1 compares the guarantees with prior algorithms.

Table 1: Sample complexity for an ε-approximate NE, measured by *Nash gap*; the four right-hand columns are decentralized, projection-free, one-sample, identity-dependent.

Algorithm	**Sample Complexity**	**Nash Regret** (NG)	**Dec.**	**PF**	**1S**	**Id.-dep.**
Nash-UCB I (Cui et al., 2022)	$\tilde{\mathcal{O}}(nF^2/\varepsilon^2)$	$\tilde{\mathcal{O}}(F\sqrt{nT})$	No	No	Yes	No
Nash-UCB II (Cui et al., 2022)	$\tilde{\mathcal{O}}(n^2F^3/\varepsilon^2)$	$\tilde{\mathcal{O}}(nF^{3/2}\sqrt{T})$	No	No	Yes	No
FW+Expl. II (Cui et al., 2022)	$\tilde{\mathcal{O}}(n^{12}F^{12}/\varepsilon^6)$	$\mathcal{O}(n^2F^2T^{5/6})$	Yes	Yes	No	No
D1S-FW (Cohen & Agmon, 2025a)	$\tilde{\mathcal{O}}(n^8/\varepsilon^2)$	$\tilde{\mathcal{O}}(n^4\sqrt{T})$	Yes	Yes	Yes	No
MF-OMD (Daskalakis et al., 2021)[§]	$\tilde{\mathcal{O}}(1/\varepsilon^2)$	$\tilde{\mathcal{O}}(\sqrt{T})$ (ext.)[§]	No	No	No	No
IndepPG (Ding et al., 2022)	$\tilde{\mathcal{O}}(n^2/\varepsilon^2)$	$\tilde{\mathcal{O}}(n\sqrt{T})$	Partial	No	No	No
OSD-FW-A, unilateral-oracle (Alg. 1, Opt. A)	Eq. (20)	$\mathcal{O}\big((n^2mv_{\max}+n\tilde{\sigma})T^{3/4}\big)$	Yes	Yes	Yes	**Yes**[†]
OSD-FW-A, pure bandit (Alg. 1, Opt. A, $\gamma > 0$)	Eq. (16)	$\mathcal{O}\big(nm^{2/3}v_{\max}T^{5/6}+n^2mv_{\max}T^{3/4}\big)$	Yes	Yes	Yes	**Yes**[†]
OSD-FW-B (Alg. 1, Opt. B, aux.-draw)	Eq. (19)	$\mathcal{O}\big((mv_{\max}+n\tilde{\sigma}+n\sigma_H)T^{2/3}\big)$	Yes[‡]	Yes	Yes	**Yes**[†]
Anon. FVG – Opt. A[¶]	Eq. (22)	$\mathcal{O}(\mathrm{poly}(n,m)\,v_{\max}T^{3/4})$	Yes	Yes	Yes	No
Anon. FVG – Opt. B[¶]	Eq. (23)	$\mathcal{O}(\mathrm{poly}(n,m)\,v_{\max}T^{2/3})$	Yes	Yes	Yes	No

F = facilities. [†]No concavity assumed; cost/variance $\mathrm{poly}(n,m)$. [‡]Opt. B needs joint-action observability and public announcements (Section 3.3.2). [¶]Conditional on Assumption 2. [§]External regret. Notes: Appendix A.4.

2 Related work

Coalition-formation and task-allocation models include hedonic games Dreze & Greenberg (1980); Aziz & Savani (2016), project games Bilò et al. (2023), structural task-allocation games Valizadeh et al. (2025a), and group activity selection problems Darmann et al. (2012). Online and dynamic versions appear in Flammini et al. (2021); Bullinger & Romen (2025; 2024); Boehmer et al. (2023); Brandt et al. (2022); Valizadeh et al. (2025b). The closest learning models are the additively separa-

ble hedonic games of Cohen & Agmon (2025a;b). Their pairwise linear structure permits efficient learning but excludes the identity-dependent coalition effects studied here.

Work on learning in potential games covers no-regret dynamics Daskalakis et al. (2021); Chen & Peng (2020), fictitious play Daskalakis & Pan (2014), and bandit feedback Heliou et al. (2017). Non-asymptotic Nash-regret guarantees have also been studied Liu et al. (2021), especially in multi-agent reinforcement learning Ding et al. (2022); Mao et al. (2022). Many of these methods require full information, central coordination, or sample complexity proportional to the action-space size. Nash-UCB Cui et al. (2022) and UCB-NS Cohen & Agmon (2025b), for example, maintain estimates centrally and do not apply to the fully decentralized model considered here.

Frank–Wolfe (conditional-gradient) methods Frank & Wolfe (1956); Jaggi (2013) avoid projections, while stochastic variants Hazan & Kale (2012); Hazan & Luo (2016); Mokhtari et al. (2018; 2020) reduce sample requirements. The one-sample method of Zhang et al. (2020) achieves $\mathcal{O}(\sqrt{T})$ rates for *convex* objectives with decaying step sizes and an $\mathcal{O}(\varepsilon^{-3})$ Frank–Wolfe-gap guarantee for a uniformly random iterate of a smooth *non-convex* objective with $\eta_t = T^{-2/3}$. The latter uses a Hessian–vector drift correction along the path between consecutive iterates. Variant B adopts this template and its $T^{-1/3}$ exponent (Section 3.4). The construction of Zhang et al. (2020), however, assumes a single optimizer that holds the increment $x^t - x^{t-1}$ and can sample along the interpolated path. In a game, the increment is the joint policy change, which no individual agent knows.

For decentralized games, Cui et al. (2022) studied mini-batch Frank–Wolfe for congestion games, and Cohen & Agmon (2025a) developed a one-sample method for additively separable hedonic games. Both use compact problem-specific structure: load counts or pairwise linearity. We consider heterogeneous fair value games with arbitrary identity-dependent coalition values. The sampled drift estimator in Section 3.3.2 has variance and per-round cost bounded by explicit polynomials in n and m (Proposition 2, Table 7). Table 3 in Appendix A.1 states the distinction precisely. Earlier decentralized methods use compact sufficient statistics; identity-dependent values admit no such statistic; and the centralized method of Zhang et al. (2020) requires an increment held by no individual agent.

3 Problem formulation

We study task-based *fair value games* Augustine et al. (2015) with a finite set of agents $N = \{1, \dots, n\}$ and tasks $M = \{j_1, \dots, j_m\}$. Each agent i selects exactly one task from $S_i \subseteq M$, and $N_j = \{i : j \in S_i\}$ is the set of agents able to perform task j. We assume $S_i \neq \emptyset$ for all i and $N_j \neq \emptyset$ for all j. For each task $j \in M$, let $v_j : 2^{N_j} \to \mathbb{R}_+ \cup \{0\}$ be a *monotone submodular* value function[1] with $v_j(\emptyset) = 0$. The function $v_j(C)$ represents the productivity of coalition C when assigned to task j. [2] For any feasible coalition $C \subseteq N_j$ and agent $i \in C$, the *marginal contribution* of agent i to task j within coalition C is $\delta_{i,j}(C) = v_j(C) - v_j(C \setminus \{i\})$.

Let $s_i \in S_i$ denote the *strategy* (task) chosen by agent i. A *strategy profile* is a vector $\mathbf{s} = (s_1, \dots, s_n)$, and the set of all strategy profiles is $\mathcal{S} = \prod_{i \in N} S_i$; we write $\mathbf{s}_{-i} = (s_k)_{k \neq i}$ and $(j, \mathbf{s}_{-i})$ for the profile obtained by replacing s_i with j.

Definition 1. *A* Fair Value Game *(FVG) is* $\mathcal{G} = \langle N, M, \mathcal{S}, (u_i)_{i \in N} \rangle$ *where* $v_{\max} = \max_{j,C} v_j(C)$ *and the utility of agent* i *under profile* $\mathbf{s}$ *is her marginal contribution to her chosen task:*

$$u_i(\mathbf{s}) = v_{s_i}\big(C_{s_i}(\mathbf{s})\big) - v_{s_i}\big(C_{s_i}(\mathbf{s}) \setminus \{i\}\big) = \delta_{i,s_i}\big(C_{s_i}(\mathbf{s})\big) \tag{1}$$

where $C_j(\mathbf{s}) = \{i \in N \mid s_i = j\}$ *is the coalition at task* j.

The policy space of agent i is $\Delta_i := \Delta(S_i) = \{\pi_i \in [0,1]^{|S_i|} : \sum_{j \in S_i} \pi_i(j) = 1\}$, and the joint space is $\boldsymbol{\Delta} = \prod_i \Delta_i$. A joint mixed strategy $\boldsymbol{\pi}$ induces independent samples $s_i \sim \pi_i$. The expected utility is $U_i(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{s} \sim \boldsymbol{\pi}}[u_i(\mathbf{s})]$, and agent i's *best-response utility* is $U_i^\star(\boldsymbol{\pi}_{-i}) = \max_{\mu_i \in \Delta(S_i)} U_i(\boldsymbol{\pi}_{-i}, \mu_i)$. For any $\varepsilon \geq 0$, a joint strategy π is an *ε-approximate Nash equilibrium*

[1] Submodularity (diminishing returns): $v(A \cup \{x\}) - v(A) \geq v(B \cup \{x\}) - v(B)$ for $A \subseteq B$ and $x \notin B$. Monotonicity: $v(A) \leq v(B)$ for $A \subseteq B$.

[2] Coalition values are task-specific: the same coalition may perform differently on different tasks, so outcomes depend on the task, coalition membership, and coalition size.

if $\max_{i\in N}\big(U_i^\star(\boldsymbol{\pi}_{-i}) - U_i(\boldsymbol{\pi})\big) \le \varepsilon$. The quantity $\mathsf{NG}(\pi) := \max_{i\in N}(U_i^\star(\boldsymbol{\pi}_{-i}) - U_i(\boldsymbol{\pi}))$ is the *Nash gap* of π.

Interpretation. The potential Φ records the change caused by every unilateral deviation. Consequently, an agent that improves her own payoff also increases Φ by the same amount. This identity connects local feedback to the Frank–Wolfe direction and, ultimately, to the Nash gap.

Lemma 1. *Every FVG is an exact potential game with potential*

$$\Phi(\mathbf{s}) \;=\; \sum_{j\in M} v_j\big(C_j(\mathbf{s})\big) \tag{2}$$

Proof sketch. When agent i switches tasks, only the two touched coalitions change; the induced change in Φ equals exactly the change in i's marginal-contribution utility, which is the exact-potential condition. Full proof in Appendix D.

Note that $\Phi(\mathbf{s}) = \sum_j v_j(C_j(\mathbf{s}))$ is the total task value, *not* the social welfare $\sum_i u_i(\mathbf{s})$ in general; the two coincide only for additive value functions. This implies the range of $\bar{\Phi}$ on $\boldsymbol{\Delta}$ is bounded by $mv_{\max}$ (see Appendix B.1).

3.1 Decentralized online learning model

We consider a fully decentralized repeated setting. At each round $t = 1, \dots, T$, each agent i independently samples $s_i \sim \pi_i^t$. Stochastic values $v_j^t(C) \sim \mathcal{V}_{j,C}$ with mean $\bar{v}_j(C) \in [0, v_{\max}]$ are realized, and each agent updates using only her own local history. There is no central coordinator and no sharing of statistics, policies, or private estimates between agents; what agents observe of the environment differs by feedback model, as we make precise below.

Feedback models. We consider three regimes, ordered by the information available to each agent. Under *bandit* feedback (Variant A), agent i observes only her own realized utility $u_i^t(\mathbf{s})$, a single scalar with no counterfactual information. This is the game's native payment rule, as in VCG/Shapley-style payments or difference rewards. This is the only fully local regime. Under the *unilateral-oracle* model (Variants A and B), the realized allocation $(C_j(\mathbf{s}))_{j\in M}$ is also visible to all agents (*joint-action observability*), and the value oracle may be queried at the $|S_i| \le m$ coalitions obtained by moving agent i alone. Under the *auxiliary-draw oracle* model (Variant B), each agent also announces two actions per round (Section 3.3.2), and the oracle may be queried at coalitions induced by profiles assembled from those announcements. These profiles may differ from the played allocation in many coordinates, although their number is $O(mn)$. The oracle is a value oracle on sets (Nemhauser et al., 1978). In platform-mediated allocation it is the platform's own scoring model $\hat{v}_j(\cdot)$, and querying it at a hypothetical team is a model evaluation, not a re-deployment. Only the bandit regime is fully local (Appendix A.2).

Nash regret. Following Ding et al. (2022); Cui et al. (2022); Cohen & Agmon (2025a), we measure performance via $\mathcal{R}^T = \sum_{t=1}^T \mathsf{NG}(\boldsymbol{\pi}^t)$. Sublinear Nash regret $\mathcal{R}^T = o(T)$ implies that the expected Nash gap of a uniformly random output iterate converges to zero: $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] = \mathcal{R}^T/T \to 0$. Since $\mathsf{NG}(\boldsymbol{\pi}) \le \sum_i (U_i^\star(\boldsymbol{\pi}_{-i}) - U_i(\boldsymbol{\pi}))$, our bounds also control the Nikaido–Isoda function Nikaidô & Isoda (1955).

3.2 Gradient alignment and the Frank–Wolfe gap

Define the multilinear extension $\bar{\Phi} : \boldsymbol{\Delta} \to \mathbb{R}$ by

$$\bar{\Phi}(\boldsymbol{\pi}) = \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\pi}}[\Phi(\mathbf{s})] = \sum_{j\in M} \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\pi}}[\bar{v}_j(C_j(\mathbf{s}))]$$

In a general FVG, $\Phi(\mathbf{s})$ differs from $\sum_i u_i(\mathbf{s})$ because the potential is not social welfare. The exact-potential property nevertheless yields the following gradient identity, the decentralized analogue of Lemma 1. Each agent's local reward gradient equals $\nabla_{\pi_i}\bar{\Phi}$ up to a constant that the Frank–Wolfe linear oracle cancels.

Lemma 2. *For every $\boldsymbol{\pi} \in \boldsymbol{\Delta}$, every $i \in N$, and every $j \in S_i$,*

$$\frac{\partial \bar{\Phi}(\boldsymbol{\pi})}{\partial \pi_i(j)} = \mathbb{E}_{\mathbf{s}_{-i} \sim \boldsymbol{\pi}_{-i}}[u_i(j, \mathbf{s}_{-i})] + c_i(\boldsymbol{\pi}_{-i}), \tag{3}$$

where $c_i(\boldsymbol{\pi}_{-i})$ is a constant independent of j. Equivalently, $\nabla_{\pi_i} \bar{\Phi}(\boldsymbol{\pi}) = \nabla_{\pi_i} U_i(\boldsymbol{\pi}) + c_i(\boldsymbol{\pi}_{-i})\mathbf{1}$.

The constant shift cancels exactly. With $\mathsf{P}_i := I - \frac{1}{|S_i|}\mathbf{1}\mathbf{1}^\top$ and $\mathsf{P} = \mathrm{diag}(\mathsf{P}_i)$, every inner product in the Frank–Wolfe analysis satisfies $\langle g, \mu_i - \pi_i \rangle = \langle \mathsf{P}_i g, \mu_i - \pi_i \rangle$ because $\mu_i - \pi_i \perp \mathbf{1}$, and $\mathsf{P}_i(c_i\mathbf{1}) = \mathbf{0}$. All errors below are therefore projected errors $e_i^t := \mathsf{P}_i(\mathbf{d}_i^t - \nabla_{\pi_i} \bar{\Phi}(\boldsymbol{\pi}^t))$, with effective variance $\tilde{\sigma}^2$ and no additive penalty (Appendix B.2).

Frank–Wolfe gap. The Frank–Wolfe gap measures the total gain agents could achieve by jointly deviating to their best responses, and equals zero if and only if the current profile is a Nash equilibrium. For $\boldsymbol{\pi} \in \boldsymbol{\Delta}$, define

$$\mathsf{FW}(\boldsymbol{\pi}) = \max_{\boldsymbol{\pi}' \in \boldsymbol{\Delta}} \langle \nabla \bar{\Phi}(\boldsymbol{\pi}), \boldsymbol{\pi}' - \boldsymbol{\pi} \rangle = \sum_i \Big[\max_{j \in S_i} \tfrac{\partial \bar{\Phi}}{\partial \pi_i(j)} - \sum_j \pi_i(j) \tfrac{\partial \bar{\Phi}}{\partial \pi_i(j)} \Big] \tag{4}$$

The second equality uses the fact that $\boldsymbol{\Delta}$ is a product of simplices, so agents' optimizations decouple.

Interpretation. For an FVG, the Frank–Wolfe gap equals total exploitability $\sum_i (U_i^\star - U_i)$. Its average therefore controls Nash regret, and a zero gap characterizes a Nash equilibrium. The general upper bound for potential games is due to Dong et al. (2024). The next lemma proves equality for FVGs and shows that the agent-specific constant in Lemma 2 cancels.

Lemma 3. *For every $\boldsymbol{\pi} \in \boldsymbol{\Delta}$, $\mathsf{FW}(\boldsymbol{\pi}) = \sum_{i \in N} (U_i^\star(\boldsymbol{\pi}_{-i}) - U_i(\boldsymbol{\pi}))$, and consequently $\mathsf{NG}(\boldsymbol{\pi}) \le \mathsf{FW}(\boldsymbol{\pi}) \le n \cdot \mathsf{NG}(\boldsymbol{\pi})$.*

The Frank–Wolfe update moves π_i toward $\arg\max_{\mu_i \in \Delta(S_i)} \langle \nabla_{\pi_i} \bar{\Phi}(\boldsymbol{\pi}^t), \mu_i \rangle$, which is the direction of agent i's best response. It is the decentralized analogue of a centralized better-response step.

Corollary 1. *A joint policy $\boldsymbol{\pi}^\star$ satisfies $\mathsf{FW}(\boldsymbol{\pi}^\star) = 0$ if and only if it is a Nash equilibrium of the mean FVG. This holds regardless of whether $\bar{\Phi}$ is concave.*

Lemma 4. *Under submodular values $v_j \in [0, v_{\max}]$, the multilinear extension $\bar{\Phi}$ is L-smooth with $L = 2(n-1)\, m\, v_{\max}$.*

Proposition 1. *Let $H = \nabla^2 \bar{\Phi}$ on the ambient box, let $\mathcal{T} = \bigoplus_i \{w_i \in \mathbb{R}^{S_i} : \langle \mathbf{1}, w_i \rangle = 0\}$ be the tangent space of $\boldsymbol{\Delta}$ and P the projection onto it. Then (a) $H_{(i,j),(k,\ell)} = 0$ unless $j = \ell$ and $i \neq k$, so H has an identically zero diagonal; and (b) $\bar{\Phi}$ is concave on $\boldsymbol{\Delta}$* iff *it is affine there,* iff $\mathsf{P}H\mathsf{P} = 0$. *The condition is on H restricted to* feasible *directions: $H \equiv 0$ is sufficient but not necessary, since an agent with $|S_i| = 1$ contributes no feasible direction. Under submodularity, if two agents share two tasks and either cross-partial is nonzero, then $\bar{\Phi}$ is not concave, as occurs generically. Appendix A.5 gives the remaining parts and a counterexample to the ambient formulation.*

3.3 OSD-FW: ONE-SAMPLE DECENTRALIZED FRANK–WOLFE

`OSD-FW` requires no coordinator and transmits no policies or private estimates. Its per-agent communication is $O(\log m)$ bits per round. This is a coordination and communication property, not a privacy guarantee; Remark 5 shows that the transcript reveals the policies. The Frank–Wolfe linear oracle costs $O(|S_i|)$ and avoids simplex projections. Each update uses one stochastic sample, which matters when a sample corresponds to a robot deployment or paid task. Sublinear Nash regret guarantees that a uniformly selected iterate approaches equilibrium in expectation.

3.3.1 STOCHASTIC GRADIENT ESTIMATORS

Each agent i constructs a stochastic estimator $\widehat{\nabla}_{\pi_i} \bar{\Phi}(\boldsymbol{\pi}^t)$ from her own feedback.

Semi-bandit estimator. Given $\mathbf{s} \sim \boldsymbol{\pi}^t$, let $C_j(j, \mathbf{s}_{-i})$ be the counterfactual coalition at task j if agent i plays j. The estimator is

$$\big[\widehat{\nabla}_{\pi_i} \bar{\Phi}(\boldsymbol{\pi}^t)\big]_j = v_j^t\big(C_j(j, \mathbf{s}_{-i})\big) - v_j^t\big(C_j(j, \mathbf{s}_{-i}) \setminus \{i\}\big), \qquad j \in S_i. \tag{5}$$

This estimates $\nabla_{\pi_i} U_i(\boldsymbol{\pi}^t)$; the additive constant $c_i\mathbf{1}$ is absorbed by the FW oracle.

Lemma 5. *With exact means,* $\mathbb{E}[\widehat{\nabla}_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) \mid \mathcal{F}_{t-1}] = \nabla_{\pi_i}U_i(\boldsymbol{\pi}^t) = \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - c_i(\boldsymbol{\pi}^t_{-i})\mathbf{1}$*, with the constant shift having no effect on the Frank–Wolfe oracle.*

Bandit gradient estimator. Under pure bandit feedback the gradient must be importance-weighted, and the weights must remain bounded. The Frank–Wolfe update drives $\min_j \pi_i^t(j)$ to zero like $(1-\eta)^t$, so both $1/\pi_i^t(s_i)$ and the estimator variance would grow without bound. Agent i therefore *plays* the explored policy $\tilde{\pi}_i^t := (1-\gamma)\pi_i^t + \gamma\,\mathrm{Unif}(S_i)$ and forms the inverse-propensity estimator (18), giving $\tilde{\sigma}^2 \le m^2 v_{\max}^2/\gamma$; unbiasedness then holds at $\tilde{\boldsymbol{\pi}}^t$, not $\boldsymbol{\pi}^t$, which is why Corollary 2 runs the argument at the played policy. Under unilateral-oracle feedback no exploration is needed and $\gamma = 0$ (Appendix A.3).

3.3.2 Frank–Wolfe update

Given a gradient estimate $\mathbf{d}_i^t$, each agent solves a linear oracle $\mu_i^t \in \arg\max_{\mu_i \in \Delta(S_i)} \langle \mathbf{d}_i^t, \mu_i \rangle$ in $\mathcal{O}(|S_i|)$ time, then updates $\pi_i^{t+1} = (1-\eta_t)\pi_i^t + \eta_t \mu_i^t \in \Delta(S_i)$ without projection.

Biased momentum (Variant A). Following Mokhtari et al. (2018; 2020):

$$\mathbf{d}_i^t = (1-\rho_t)\mathbf{d}_i^{t-1} + \rho_t \widehat{\nabla}_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t), \qquad \rho_t = \tfrac{2}{(t+4)^{1/2}},\ \mathbf{d}_i^0 = \mathbf{0}. \tag{6}$$

This estimator is biased but has vanishing mean-squared error. The momentum exponent $\frac{1}{2}$ is not the $\frac{2}{3}$ used for centralized SCG Mokhtari et al. (2018): that choice is tuned to a *decaying* step size $\eta_t = 1/t$, whereas we analyse a *constant* η (Section 3.4), which changes the optimal balance between momentum lag and momentum noise. Appendix D records the trade-off.

Unbiased momentum (Variant B). To remove the momentum lag we adopt the unbiased-momentum recursion of Zhang et al. (2020),

$$\mathbf{d}_i^t = (1-\rho_t)(\mathbf{d}_i^{t-1} + \tilde{\boldsymbol{\Delta}}_i^t) + \rho_t \widehat{\nabla}_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t), \tag{7}$$

in which $\tilde{\boldsymbol{\Delta}}_i^t$ must be a conditionally unbiased estimator of the gradient increment $\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1})$. This estimator is the component that must be changed when the centralized template is applied to a game.

Limits of the centralized construction. Zhang et al. (2020) take $\tilde{\boldsymbol{\Delta}}^t = \widetilde{\nabla}^2 F(x^t(\alpha); z)(x^t - x^{t-1})$ along the interpolated path $x^t(\alpha) = x^{t-1} + \alpha(x^t - x^{t-1})$. Transcribed to a game this multiplies the *joint* increment, of which agent i holds one block, and requires a sample from $\boldsymbol{\pi}^t(\alpha)$, an interpolation of policies she cannot see. Observing an action does not reveal the distribution that generated it. Agent i therefore cannot draw the required interpolated sample, and a different construction is needed.

Auxiliary-draw protocol. Two observations make the drift correction implementable without any agent learning another's policy. First, the Frank–Wolfe update pins the *shape* of the increment:

$$\boldsymbol{\pi}_k^t - \boldsymbol{\pi}_k^{t-1} = \eta\left(\mu_k^{t-1} - \pi_k^{t-1}\right), \tag{8}$$

where η is a publicly known parameter and μ_k^{t-1} is the vertex returned by agent k's *own* linear oracle at round $t-1$. Second, every quantity the estimator needs from agent k is an expectation under π_k^{t-1} or μ_k^{t-1}. That expectation can be estimated from one sample, which agent k is able to draw. Accordingly, at the start of round t each agent k announces two actions,

$$b_k^t := \text{the task supported by } \mu_k^{t-1},\ a_k^t \sim \pi_k^{t-1} \text{ drawn afresh, independently of everything else.} \tag{9}$$

Both are single elements of S_k, of exactly the same type as a played action. Agent i then draws privately $\alpha^t \sim \mathrm{Unif}[0,1]$ and, independently across $k \neq i$, a coin $c_k^t \sim \mathrm{Bern}(\alpha^t\eta)$, and assembles

$$\mathbf{s}_k'^t := b_k^t \text{ if } c_k^t = 1, \qquad \mathbf{s}_k'^t := a_k^t \text{ otherwise}, \qquad k \neq i. \tag{10}$$

The construction is exact. Substituting (8) into $\pi_k^t(\alpha) := \pi_k^{t-1} + \alpha(\pi_k^t - \pi_k^{t-1})$ gives $\pi_k^t(\alpha) = (1-\alpha\eta)\,\pi_k^{t-1} + \alpha\eta\,\mu_k^{t-1}$, so $\mathbf{s}_k'^t$ is a draw from $\pi_k^t(\alpha)$ and, by independence across agents, $\mathbf{s}'^t \sim \boldsymbol{\pi}^t(\alpha)$ on the coordinates that matter. Agent i has therefore produced a sample from the interpolated joint policy using only the announced actions.

Algorithm 1 OSD−FW for Fair Value Games

Require: Horizon T; variant A (biased) or B (unbiased). *Parameters:* A: $\eta = T^{-3/4}$, $\rho_t = 2/(t{+}4)^{1/2}$, $\gamma = 0$ (unilateral) or $\Theta(m^{2/3}T^{-1/6})$ (bandit); B: $\eta = T^{-2/3}$, $\rho_t = (t{+}1)^{-2/3}$, $\gamma = 0$

Each agent i initialises π_i^1 uniform on S_i; $\mathbf{d}_i^0 = \mathbf{0}$; $\mu_i^0 \leftarrow$ any vertex of $\Delta(S_i)$
for $t = 1, \dots, T$ **do**
  Variant B only, announcements: each agent k publishes $b_k^t \leftarrow$ the task supported by μ_k^{t-1}, and a fresh independent draw $a_k^t \sim \pi_k^{t-1}$ (9)
  Each agent i forms $\tilde{\pi}_i^t = (1-\gamma)\pi_i^t + \gamma\,\mathrm{Unif}(S_i)$ and **plays** $s_i \sim \tilde{\pi}_i^t$; observes her feedback
  Compute $\widehat{\nabla}_{\pi_i}\bar{\Phi}$ via (5) (unilateral-oracle) or (18) (bandit, Variant A only)
  Variant A: $\mathbf{d}_i^t = (1-\rho_t)\mathbf{d}_i^{t-1} + \rho_t \widehat{\nabla}_{\pi_i}\bar{\Phi}$
  Variant B: draw $\alpha^t \sim \mathrm{Unif}[0,1]$ and coins $c_k^t \sim \mathrm{Bern}(\alpha^t\eta)$; assemble $\mathbf{s}'^t$ by (10); compute $\tilde{\boldsymbol{\Delta}}_i^t$ by (11); $\mathbf{d}_i^t = (1-\rho_t)(\mathbf{d}_i^{t-1} + \tilde{\boldsymbol{\Delta}}_i^t) + \rho_t \widehat{\nabla}_{\pi_i}\bar{\Phi}$
  $\mu_i^t \in \arg\max_{\mu_i \in \Delta(S_i)} \langle \mathbf{d}_i^t, \mu_i \rangle$; $\pi_i^{t+1} = (1-\eta)\pi_i^t + \eta\,\mu_i^t$
end for
Sample $\tau \sim \mathrm{Unif}\{1, \dots, T\}$
return $\tilde{\boldsymbol{\pi}}^\tau$

Sampled drift correction. Agent i forms, for $j \in S_i$,

$$\big[\tilde{\boldsymbol{\Delta}}_i^t\big]_j = \eta \sum_{k \neq i} \Big[\Phi\big(j, b_k^t, \mathbf{s}'^t_{-\{i,k\}}\big) - \Phi\big(j, a_k^t, \mathbf{s}'^t_{-\{i,k\}}\big) \Big]. \tag{11}$$

Each summand is a difference of potentials at two profiles differing only in agent k's action, hence reduces to two marginal contributions and costs $O(1)$ value-oracle calls; the whole vector costs $O(mn)$ calls, *fewer* than the $O(m^2 n)$ of the Jacobian–vector form it replaces. Note also that the baseline term $\Phi(j, s_k, \mathbf{s}_{-\{i,k\}})$ appearing in the Jacobian form has disappeared: it cancels identically because $\sum_\ell (\pi_k^t(\ell) - \pi_k^{t-1}(\ell)) = 0$.

Lemma 6. *Let the announcements (9) be drawn independently across agents and independently of $\mathcal{F}_{t-1}$ given $\boldsymbol{\pi}^{t-1}$, and let α^t and the coins $\{c_k^t\}_{k \neq i}$ be drawn by agent i independently of everything else. Then the estimator (11) satisfies*

$$\mathbb{E}\big[\tilde{\boldsymbol{\Delta}}_i^t \mid \mathcal{F}_{t-1}\big] = \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1}), \qquad \textit{and} \qquad \big\|\tilde{\boldsymbol{\Delta}}_i^t\big\| \le 2\eta(n-1)\sqrt{m}\, v_{\max} \ \textit{a.s.}$$

Agent i's computation reads only $\{a_k^t, b_k^t\}_{k \neq i}$, η, her own randomness and oracle responses.

3.4 Convergence analysis

We state the assumptions and main convergence theorems.

Assumption 1. *(i) $\boldsymbol{\Delta}$ has Euclidean diameter $D \le \sqrt{2n}$. (ii) The projected gradient estimator has per-agent variance $\le \tilde{\sigma}^2$ around $\nabla_{\pi_i} U_i$, and for Variant B the projected drift estimator (11) satisfies*

$$\mathbb{E}\big[\|\mathsf{P}_i(\tilde{\boldsymbol{\Delta}}_i^t - (\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1})))\|^2 \mid \mathcal{F}_{t-1}\big] \le \eta^2 \sigma_H^2. \tag{12}$$

(iii) $v_j^t(C) \in [0, v_{\max}]$ almost surely.

Both bounds are per agent; the proofs use the stacked error over all n agents, so they enter as $n\eta^2\sigma_H^2$ and $n\tilde{\sigma}^2$. Explicit polynomial bounds are $\tilde{\sigma}^2 \le m v_{\max}^2$ (unilateral-oracle), $\tilde{\sigma}^2 \le m^2 v_{\max}^2/\gamma$ (bandit), and $\sigma_H^2 \le 4(n-1)^2 m v_{\max}^2$ (Proposition 2, Appendix A.6). *No concavity assumption appears*: by Proposition 1 none is available.

3.4.1 Biased variant (Variant A)

Theorem 1. *Under Assumption 1, with $\eta_t = T^{-3/4}$ and $\rho_t = 2/(t+4)^{1/2}$, Algorithm 1 Variant A under unilateral-oracle feedback ($\gamma = 0$) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)] \ \le \ \frac{m v_{\max}}{T^{1/4}} + \frac{5D\sqrt{Q}}{T^{1/4}} + \frac{LD^2}{2\,T^{3/4}}, \tag{13}$$

with Q as in Lemma 7. Substituting $L = 2(n-1)mv_{\max}$ and $D^2 \le 2n$ gives $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] \le \frac{1}{T}\sum_t \mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)] = \mathcal{O}((n^2 m v_{\max} + n\tilde{\sigma})T^{-1/4})$, i.e. $\mathcal{R}^T = \mathcal{O}((n^2 m v_{\max} + n\tilde{\sigma})T^{3/4})$.

Under *pure bandit* feedback, running Variant A with exploration weight γ and regarding the played sequence $\tilde{\boldsymbol{\pi}}^t$ as Frank–Wolfe iterates on the shrunken product $\boldsymbol{\Delta}^\gamma$ removes the bias entirely and gives $\mathcal{R}^T = \mathcal{O}(n^2 m v_{\max} T^{3/4} + n m^{2/3} v_{\max} T^{5/6})$ at $\gamma^\star = \Theta(m^{2/3}T^{-1/6})$ (Corollary 2, Appendix A.5). As noted, this exponent is superseded by Dong et al. (2024).

3.4.2 Unbiased variant (Variant B)

Theorem 2. *Under Assumption 1, with $\eta_t = T^{-2/3}$ and $\rho_t = (t+1)^{-2/3}$, Algorithm 1 Variant B satisfies*

$$\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)] \le \frac{m v_{\max}}{T^{1/3}} + \frac{5D\sqrt{Q_B}}{T^{1/3}} + \frac{LD^2}{2T^{2/3}}, \quad Q_B = 3n\tilde{\sigma}^2 + 6n\sigma_H^2 + 2nmv_{\max}^2, \tag{14}$$

so $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] = \mathcal{O}((m v_{\max} + n\tilde{\sigma} + n\sigma_H)T^{-1/3})$ and $\mathcal{R}^T = \mathcal{O}((m v_{\max} + n\tilde{\sigma} + n\sigma_H)T^{2/3})$. To reach $\mathbb{E}[\mathsf{NG}] \le \varepsilon$ this needs $T = \mathcal{O}(n^6 m^{3/2} v_{\max}^3/\varepsilon^3)$ rounds (Corollary 3). As an optimization statement this is Theorem 2 of Zhang et al. (2020) instantiated at $\bar{\Phi}$ with estimator (11); what is specific to FVGs is that by Lemma 3 the left-hand side is Nash regret.

4 Experimental analysis

All results come from one reference implementation of Algorithm 1, with the Nash gap, welfare and potential computed *exactly* on the policy played in every round by summation over coalition masks ($\Theta(nm2^n)$). The theorems bound $\mathcal{R}^T/T = \mathbb{E}[\mathsf{NG}(\tilde{\boldsymbol{\pi}}^\tau)]$ for uniform τ, so that is the primary metric and the final-iterate gap is reported separately. Baselines are IndepPG, IndepMD, UCB-MC, ε-Greedy, FW-Exploration, an exact-gradient MF-OMD reference, and a faithful transcription of Dong et al. (2024). Environments, seeds and complete tables are in Appendix E.

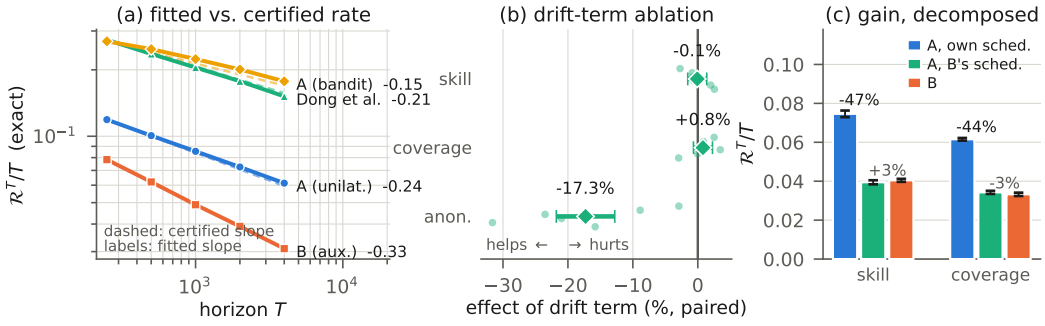

Figure 1: Empirical rates and drift-correction ablations. (a) Average Nash gap with fitted and certified slopes. (b) Paired effect of the drift term. (c) Decomposition of the Variant A–B difference into schedule and estimator effects.

Matched oracle budget. An equal-T comparison ignores the large difference in per-round oracle use ($O(mn)$ for Variant B and $O(1)$ for the bandit methods), so Table 6 fixes a budget of $\approx 1.98 \times 10^5$ calls and gives each method the horizon its own rate affords; cheap methods run 20–22× longer. On $\mathcal{R}^T/T$, `OSD-FW-B` is *fourth* (0.0342 ± 0.0011), behind UCB-MC (0.0047), IndepPG (0.0075) and ε-Greedy (0.0130); on the final iterate it is second (0.00033 ± 0.00020), behind UCB-MC (0.00000). UCB-MC is decentralized (Appendix A.7); B has the best final gap only among the Frank–Wolfe methods. Dong et al. (2024) place ninth (0.0860), ahead of our pure-bandit variant (0.1138), consistent with their better exponent. There is thus no Nash-regret advantage here; on cumulative exploitability the cheaper methods benefit more from the longer horizon the budget affords, and the final-iterate quantity on which B leads is one our analysis does not cover.

Table 2: Matched value-oracle budget on an identity-dependent FVG (n=6, m=4, 9 runs). Each method runs until it reaches the same call budget (spread 11.8%). Because the step size depends on T, we solve the fixed point $T \leftarrow T \cdot \mathrm{budget}/\mathrm{spend}(T)$. Metrics are exact over each method's horizon; $\pm$ is one standard error. IndepMD (0.0400) and FW-Exploration (0.0772) are reported in Appendix A.7.

Method	Feedback	T	Calls	$\mathcal{R}^T/T \downarrow$	Final gap $\downarrow$
UCB-MC	Bandit	32,933	197,598	$\mathbf{0.0047 \pm 0.0005}$	$\mathbf{0.00000 \pm 0.00000}$
IndepPG	Unilateral	4,117	197,616	0.0075 ± 0.0004	0.00102 ± 0.00063
ε-Greedy	Bandit	32,933	197,598	0.0130 ± 0.0004	0.01171 ± 0.00037
`OSD-FW`-B (sampled)	Aux.-draw	2,060	174,288	0.0342 ± 0.0011	0.00033 ± 0.00020
`OSD-FW`-A	Unilateral	4,117	197,616	0.0530 ± 0.0006	0.00099 ± 0.00043
`OSD-FW`-B (exact JVP)	Aux.-draw	375	197,520	0.0626 ± 0.0025	0.00145 ± 0.00058
Dong et al. (2024)	Bandit	32,933	197,598	0.0860 ± 0.0030	0.04532 ± 0.00588
`OSD-FW`-A (bandit)	Bandit	32,933	197,598	0.1138 ± 0.0024	0.08859 ± 0.00173

Empirical rates. Regressing $\log(\mathcal{R}^T/T)$ on $\log T$ over $T \in \{250, \ldots, 4000\}$ gives fitted slopes -0.239, -0.335, -0.151 and -0.211 against certified $-1/4$, $-1/3$, $-1/6$ and, for Dong et al. (2024), $-1/5$, with $R^2 \geq 0.994$ (Figure 1a). Fitting the *final-iterate* gap instead gives roughly -1.4 and -1.0 for the two oracle-feedback variants, which would suggest the bounds are loose; that is an artifact of fitting a quantity the theorems do not bound. The slopes are consistent with the certified decay; tightness would need a matching lower bound. The ranking is unchanged at equal T and on a *non-monotone* environment outside Definition 1 (Appendix A.7).

Drift-correction ablation. The rate comparison does not isolate the drift correction, so we ablate it directly. Holding the schedule, instances and seeds fixed and removing only the drift term from Variant B leaves $\mathcal{R}^T/T$ statistically unchanged on both identity-dependent families: -0.1% (95% CI $[-1.6, +1.4]$) on SKILL and $+0.8\%$ ($[-0.7, +2.3]$) on COVERAGE. It improves the metric by 17.3% ($[-21.8, -12.8]$) only on the anonymous family (Figure 1b). The difference between Variants A and B at equal T comes mainly from their schedules: running Variant A on Variant B's step-size schedule closes 47% of the gap on SKILL and 44% on COVERAGE, and exchanging the estimator then moves the remainder by only $+2.6\%$ and -2.9% (Figure 1c). The correction supports the $T^{-1/3}$ proof but has no measurable effect on these identity-dependent instances.

5 Conclusion

We introduced `OSD-FW`, a projection-free one-sample Frank–Wolfe method for decentralized online learning in fair value games. Its new component is a sampled drift correction, which replaces the joint policy increment no agent holds by two actions per agent per round, each drawn and announced by its owner, at $O(mn)$ oracle calls in place of $O(m^2 n)$. Because the Frank–Wolfe gap *equals* total exploitability in an FVG, the resulting stationarity bounds are Nash-regret bounds — $\mathcal{O}(T^{3/4})$ under unilateral-oracle feedback, $\mathcal{O}(T^{2/3})$ with the correction, $\mathcal{O}(T^{5/6})$ under pure bandit feedback — and the fitted exponents match the certified ones to within 0.016. A controlled ablation then locates what the correction is actually worth: it is inert on the identity-dependent instances the model is built around, its apparent gain over the biased variant is a step-size effect, and at a matched oracle budget it does not attain the lowest Nash regret. It pays only where the gradient path is incoherent, which these families are not. What a drift correction buys is a property of the dynamics, not of the value class.

The sharpest limitation is that our analysis cannot express this. Its lag term depends on the instance only through the worst-case smoothness constant, so it credits the correction uniformly where measurement finds it worth between nothing and 16%. Beyond that, the exponents describe the constant-step family analysed here, the pure-bandit rate is superseded by Dong et al. (2024), the final-iterate behaviour on which Variant B compares best lies outside the theory, and the anonymous rows of Table 1 remain conditional on Assumption 2. Appendix C turns each of these into a concrete direction, together with the path functional that would make a bound sensitive to the effect in Figure 1b.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and assist with implementing parts of the experiments. We did not use these tools to develop the novelty of the work, discover literature, draft sections of the paper, generate synthetic datasets, or formulate or prove mathematical claims. All AI-assisted text and code were reviewed, tested, and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

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

Haris Aziz and Rahul Savani. Hedonic games. 2016.

Andrew An Bian, Baharan Mirzasoleiman, Joachim M. Buhmann, and Andreas Krause. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 111–120, 2017.

Jakob Bichler, Andreu Matoses Gimenez, and Javier Alonso-Mora. Sadcher: Scheduling using attention-based dynamic coalitions of heterogeneous robots in real-time. *arXiv preprint arXiv:2510.14851*, 2025.

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

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

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

Felix Brandt, Martin Bullinger, and Leo Tappe. Single-agent dynamics in additively separable hedonic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4867–4874, 2022.

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

Martin Bullinger and René Romen. Online coalition formation under random arrival or coalition dissolution. *ACM Transactions on Algorithms*, 22(1):1–43, 2025.

Yishuai Cai, Xinglin Chen, Zhongxuan Cai, Yunxin Mao, Minglong Li, Wenjing Yang, and Ji Wang. Mrbtp: Efficient multi-robot behavior tree planning and collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14548–14557, 2025.

Gruia Calinescu, Chandra Chekuri, Martin Pal, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

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

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

Saar Cohen and Noa Agmon. Online learning of partitions in additively separable hedonic games. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2722–2730, 2024.

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

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

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

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

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

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

Jing Dong, Baoxiang Wang, and Yaoliang Yu. Convergence to nash equilibrium and no-regret guarantee in (markov) potential games. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

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

Michele Flammini, Gianpiero Monaco, Luca Moscardelli, Mordechai Shalom, and Shmuel Zaks. On the online coalition structure generation problem. *Journal of Artificial Intelligence Research*, 72:1215–1250, 2021.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Elad Hazan and Satyen Kale. Projection-free online learning. *arXiv preprint arXiv:1206.4657*, 2012.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pp. 1263–1271. PMLR, 2016.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 15007–15049. PMLR, 2022.

Leandro Soriano Marcolino, Albert Xin Jiang, and Milind Tambe. Multi-agent team formation: Diversity beats strength? In *IJCAI*, volume 13, 2013.

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

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

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. 1955.

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

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

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

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

Jinfu Xia, Yan Zhao, Guanfeng Liu, Jiajie Xu, Min Zhang, and Kai Zheng. Profit-driven task assignment in spatial crowdsourcing. In *IJCAI*, pp. 1914–1920, 2019.

Fuhan Yan and Kai Di. Multi-robot task allocation in the environment with functional tasks. In *IJCAI*, pp. 4710–4716, 2022.

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

Shuying Zhang, Zuyao Ni, Linling Kuang, Chunxiao Jiang, and Xu Zhao. Load-aware distributed resource allocation for mf-tdma ad hoc networks: A multi-agent drl approach. *IEEE Transactions on Network Science and Engineering*, 9(6):4426–4443, 2022.

A SUPPLEMENTARY DISCUSSION

A.1 SCOPE AND RELATION TO PRIOR WORK

Table 3: Positioning of our one-sample Frank–Wolfe update against prior stochastic- and decentralized-FW methods. The novelty is the estimator needed once no compact statistic of the joint policy exists *and* no agent holds the joint iterate increment; the surrounding non-convex analysis is inherited from Zhang et al. (2020).

Method	**Statistic exploited**	**Why it fails for identity-dependent FVGs**
Cui et al. (2022) (congestion)	resource load counts $\lvert C_j\rvert$	value depends on coalition *identity*, not size
Cohen & Agmon (2025a) (hedonic)	pairwise-linear coefficients	v_j is a general submodular set function, not a sum of pairs
Zhang et al. (2020) (centralized)	the optimizer's own iterate increment $x^t - x^{t-1}$	no agent knows the *joint* policy change, nor can sample its interpolation
Ours (`OSD-FW`)	*none available*	new *sampled* drift correction from public action draws (Sec. 3.3.2)

Remark 1 (Hessian restriction to feasible directions)**.** *An earlier formulation of part (b) read "concave on $\boldsymbol{\Delta}$ iff $H \equiv 0$". That is false when some agent has a singleton feasible set, because such an agent contributes no feasible direction and her rows of H are then unconstrained. Take $n = 2$, $S_1 = \{A\}$, $S_2 = \{A, B\}$, $v_A(C) = \min\{|C|, 1\}$ and $v_B(C) = |C|$. Then $\Phi(A, A) = 1$ and $\Phi(A, B) = 2$, so with $\pi_2 = (q, 1-q)$ we get $\Phi = 2 - q$, which is affine, hence concave, on $\boldsymbol{\Delta}$. Yet the ambient cross-partial is $H_{(1,A),(2,A)} = v_A(\{1,2\}) - v_A(\{1\}) - v_A(\{2\}) + v_A(\emptyset) = 1 - 1 - 1 + 0 = -1 \neq 0$. The corrected statement is consistent with this: agent 1 has $|S_1| = 1$, so $w_1 = 0$ for every $w \in \mathcal{T}$ and $\mathsf{P}H\mathsf{P} = 0$. Nothing else in the paper changes, because every use of H in the analysis is already through P (Section 3.2); but the interpretation must be stated correctly: concavity is equivalent to the absence of cross-agent interaction* along directions the agents can actually move in*, which in the example is the statement that a game with a constrained agent is not strategically two-player at all.*

Remark 2 (Role of submodularity)**.** *Lemma 1 (exact potential game) and Corollary 1 (FW-stationary $\Leftrightarrow$ Nash) hold for every value function $v_j : 2^{N_j} \to \mathbb{R}$, because their proofs use only marginal contributions. Thus, `OSD-FW`'s best-response construction and Nash characterization allow complementarities and supermodular effects. Submodularity supplies the explicit smoothness constant $L = 2(n-1)mv_{\max}$ in Lemma 4. Without submodularity, the theorems use a larger, instance-dependent constant L'. Submodularity also appears in part (c) of Proposition 1, which no proof invokes. The convergence rates require smoothness and boundedness. We have not tested supermodular or mixed-interference instances empirically.*

Remark 3. *Algorithm 1 returns the played iterate $\tilde{\boldsymbol{\pi}}^\tau$ at a uniformly random round τ. It does not return the time average $\bar{\boldsymbol{\pi}} = \frac{1}{T}\sum_t \tilde{\boldsymbol{\pi}}^t$. Because NG is not convex in $\boldsymbol{\pi}$, Jensen's inequality does not give $\mathbb{E}[\mathsf{NG}(\bar{\boldsymbol{\pi}})] \leq \frac{1}{T}\sum_t \mathbb{E}[\mathsf{NG}(\tilde{\boldsymbol{\pi}}^t)]$; the random-iterate construction satisfies this by linearity. This matches the convention of Zhang et al. (2020). Under unilateral-oracle feedback $\gamma = 0$ and $\tilde{\boldsymbol{\pi}}^t = \boldsymbol{\pi}^t$; the distinction matters only in the pure-bandit case, where the guarantee must be stated for the policy actually played.*

Remark 4 (Non-vacuous horizon thresholds)**.** *The constants are large. Since $\mathsf{NG} \leq 2v_{\max}$ holds trivially, the Variant B bound becomes informative only when $T \gtrsim \big((mv_{\max} + n\tilde{\sigma} + n\sigma_H)/2v_{\max}\big)^3$; substituting the explicit polynomial bounds on $\tilde{\sigma}$ and σ_H gives $T = \Theta(n^6 m^{3/2})$, about 3×10^7 rounds at n=m=10, and $\Theta(n^8 m^4)$ for Variant A. Those are worst-case thresholds over all monotone submodular v_j with the stated bounds. In the experiments, Section 4 reaches a final-iterate Nash gap of 1×10^{-4} at n=8, m=5, T=2000, far inside the horizon the bound would demand. The* average *gap the bound actually governs is larger, 4.2×10^{-2}, and tracks the certified exponent closely (Figure 1a). Thus, the loose constants appear in the difference between the average and final gaps, while the empirical rate agrees closely with the bound. The discrepancy is concentrated in the smoothness term $LD^2 = O(n^2 m v_{\max})$ and in Proposition 2, which bounds the drift by treating all $n-1$ per-agent contributions in (11) as simultaneously maximal and perfectly correlated; once the*

Frank–Wolfe directions stabilize, most of those contributions are identically zero because a_k^t and b_k^t coincide. Tighter instance-dependent constants may explain this discrepancy.

A.2 Feedback requirements and assumptions

Remark 5 (Policy reconstruction from the transcript)**.** *The protocol provides no privacy guarantee. The announced vertex b_k^t is the support of μ_k^{t-1}, the step size η is a public parameter, and Algorithm 1 initialises π_k^1 uniform on the known S_k. Hence*

$$\pi_k^t = (1-\eta)\,\pi_k^{t-1} + \eta\, e_{b_k^t}, \qquad \pi_k^1 = \mathrm{Unif}(S_k),$$

is a deterministic recursion driven entirely by public data, so an observer who logs $\{b_k^r\}_{r\le t}$ reconstructs $\pi_k^1, \ldots, \pi_k^t$ exactly. Although a single action sample does not reveal a distribution, the complete transcript reveals every policy. The protocol therefore provides no privacy guarantee.

The construction reduces communication and oracle queries. Every transmitted object is a single action, so the per-agent message is $2\lceil \log_2 m \rceil$ bits per round and is independent of n. Agents transmit no real-valued policy, gradient, or momentum vector. The sampled form also costs $O(mn)$ value-oracle calls, compared with $O(m^2 n)$ for the Jacobian–vector form, because the sum over ℓ collapses onto two sampled actions.

Since the transcript determines π_k^{t-1}, an agent could in principle reconstruct it and evaluate $\sum_\ell \Phi(j,\ell,\cdot)\,\pi_k^{t-1}(\ell)$ exactly instead of sampling a_k^t. The sampled draw reduces queries by trading an exact $O(m)$-term sum for one oracle evaluation. This is where the factor m in the query count is saved, at the price of the variance σ_H^2 of Proposition 2.

A privacy-preserving version would have to suppress or randomize the announced vertices, which would require a different analysis.

Remark 6. *The results, including Corollary 1, do not assume concavity of $\bar{\Phi}$. The convergence analysis of Section 3.4 bounds the* average *Frank–Wolfe gap $\frac{1}{T}\sum_t \mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)]$ by a smoothness-plus-boundedness telescoping argument valid for an arbitrary smooth, bounded, non-concave $\bar{\Phi}$; the average-gap form matters because, by Lemma 3, it is a Nash-regret bound, whereas a min-iterate bound Lacoste-Julien (2016) would certify only a single unidentified round. Both the non-convex average-gap argument (Zhang et al., 2020) and the average-gap-to-Nash-regret step (Dong et al., 2024) are prior work and are used here with attribution.*

Feedback required by each variant

What each variant actually needs (stated precisely). We are explicit about this because the two variants sit at genuinely different points on the observability scale, and an earlier informal description of eq. (5) understated what it requires.

- **Variant A, pure bandit.** The agent observes only the scalar $u_i^t(\mathbf{s})$. This is the fully local, own-observation-only regime.
- **Variant A or B, unilateral-oracle.** The estimator (5) evaluates v_j at $C_j(j, \mathbf{s}_{-i})$ for every $j \in S_i$, i.e. at the coalition that *would* form at task j had agent i chosen j. Constructing that set requires knowing *which* agents currently occupy task j; a value oracle over sets does not by itself tell an agent which set to query. Our unilateral-oracle model therefore includes **joint-action observability**: the realized allocation is visible to all agents. Cost: $|S_i| \le m$ oracle calls per round, on *single*-deviation coalitions.
- **Variant B, drift correction.** Everything Variant A needs under unilateral-oracle feedback, plus the auxiliary-draw protocol above: the $2(n-1)$ announced actions $\{a_k^t, b_k^t\}_{k\neq i}$, the publicly known step size η, and $O(mn)$ value-oracle calls per round at profiles assembled from those announcements. **Agent i is never *sent* another agent's policy or policy increment**: her update is a function of announced actions, public parameters, her own randomness and oracle responses. This is what makes the drift correction computable at all, since the joint increment $\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}$ that the centralized construction multiplies is held by nobody. It is *not* a secrecy property, the transcript is policy-revealing (Remark 5).

Every auxiliary message contains one action, so the per-agent communication is $2\lceil \log_2 m \rceil$ bits per round. Agents transmit no policy, gradient, statistic, or private estimate. Nevertheless, the

announced vertices, public η, and prescribed initialization determine every policy sequence exactly (Remark 5). The protocol reduces communication and query cost but provides no privacy. Its observability requirements are strictly stronger than own-payoff feedback, so Variants A-unilateral and B use a different notion of decentralization from Variant A under pure bandit feedback. Table 1 marks this distinction, and Section B.6 discusses it further.

When is value-oracle feedback realistic? A platform-mediated allocation system, such as a crowdsourcing marketplace or multi-robot dispatcher, may already maintain a learned or simulated model $\hat{v}_j(\cdot)$ of team performance and publish the current allocation. In that case, a counterfactual query is a model evaluation and requires no physical redeployment. Variant B also requires each participant to publish two sampled candidate actions per round. A platform can collect these actions as intent or bid messages, although the resulting transcript reveals the policies (Remark 5). When only an agent's realized payoff is observable, Variant A supplies the slower $T^{5/6}$ pure-bandit guarantee.

WORKED EXAMPLE WITH COUNTERFACTUAL ORACLE QUERIES

Worked example (why no counterfactual coalition is physically formed). Take three skill-weighted workers, two tasks, and $v_j(C) = \big(\sum_{i\in C} W_{ij}\big)/\sqrt{|C|}$. In a round, worker 1 plays task A while task B is served by $\{2,3\}$; unilateral-oracle feedback reveals the realized coalition and value $v_B(\{2,3\})$. To form the $j{=}B$ component of (5), worker 1 needs $v_B(\{1,2,3\}) - v_B(\{2,3\})$: a single query of the identity-indexed scoring model at the set $\{1,2,3\}$, a coalition *one member away* from the one just observed. Because the oracle is a set function of identities, this handles the non-anonymous (identity-dependent) case directly. Only Variant B needs these queries; Variant A uses the single realized reward.

A.3 DETAILED STATEMENTS OF SCOPE AND CONTRIBUTIONS

DETAILED STATEMENT OF PROPOSITION 1

The statement below is the long form of Proposition 1; the main text gives the compact version.

Lemma 7. *Let $e_i^t := \mathsf{P}_i(\mathbf{d}_i^t - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t))$ and let the stacked error be $e^t = (e_1^t, \dots, e_n^t)$. Assume L-smoothness (Lemma 4), a constant step size $\eta_t \equiv \eta$, and $\mathbb{E}[\|\mathsf{P}_i(\widehat{\nabla}_{\pi_i}\bar{\Phi} - \nabla_{\pi_i}U_i)\|^2] \le \tilde{\sigma}^2$ for each i. Then the estimator (6) satisfies*

$$\mathbb{E}\big[\|e^t\|^2\big] \;\le\; \frac{Q}{(t+5)^{1/2}}, \qquad Q \;=\; 3n\tilde{\sigma}^2 + 10\,L^2D^2\eta^2T^{3/2} + 3nm\,v_{\max}^2.$$

In particular, with $\eta = T^{-3/4}$ the constant $Q = 3n\tilde{\sigma}^2 + 10L^2D^2 + 3nmv_{\max}^2$ is free of T.

Corollary 2 (Pure bandit feedback)**.** *Run Variant A with exploration weight $\gamma \in (0,1)$ and let $\boldsymbol{\Delta}^\gamma := \prod_i\{(1-\gamma)\pi_i + \gamma\,\mathrm{Unif}(S_i) : \pi_i \in \Delta(S_i)\}$ be the shrunken product of simplices. Then Theorem 1 applies* verbatim *to the played sequence $\tilde{\boldsymbol{\pi}}^t$ regarded as Frank–Wolfe iterates on $\boldsymbol{\Delta}^\gamma$, with $\tilde{\sigma} \le mv_{\max}/\sqrt{\gamma}$ and* no bias term*, and*

$$\mathcal{R}^T = \sum_t \mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \;\le\; \mathcal{O}\big(n^2mv_{\max}T^{3/4} + nmv_{\max}\gamma^{-1/2}T^{3/4}\big) + n\gamma v_{\max}T. \tag{15}$$

Optimizing the last two terms gives $\gamma^\star = \Theta(m^{2/3}T^{-1/6})$ and $\mathcal{R}^T = \mathcal{O}(n^2mv_{\max}T^{3/4} + nm^{2/3}v_{\max}T^{5/6})$, so setting $\mathcal{R}^T/T \le \varepsilon$,

$$T = \mathcal{O}\left(\frac{n^6m^4v_{\max}^6}{\varepsilon^6} \;+\; \frac{n^8m^4v_{\max}^4}{\varepsilon^4}\right). \tag{16}$$

Forced exploration is what makes the pure-bandit guarantee hold at all*: without it $1/\tilde{\pi}_i^t(s_i)$ is unbounded along the Frank–Wolfe trajectory and no variance bound is available. Its price is the gap between $T^{3/4}$ and $T^{5/6}$.*

PURE-BANDIT GRADIENT ESTIMATOR

Bandit gradient estimator (with forced exploration). Under bandit feedback, agent i observes only $u_i^t(\mathbf{s})$, so the gradient must be importance-weighted, and the importance weights must be controlled. This is not a formality: the Frank–Wolfe update $\boldsymbol{\pi}^{t+1} = (1-\eta)\boldsymbol{\pi}^t + \eta\boldsymbol{\mu}^t$ moves toward a

vertex of $\Delta(S_i)$, so $\min_j \pi_i^t(j)$ decays like $(1-\eta)^t$ and would drive $1/\pi_i^t(s_i)$, hence the estimator variance, beyond any constant. We therefore *play* the explored policy

$$\tilde{\pi}_i^t := (1-\gamma)\,\pi_i^t + \gamma\,\mathrm{Unif}(S_i), \qquad \gamma \in (0,1), \qquad s_i \sim \tilde{\pi}_i^t, \tag{17}$$

which guarantees $\tilde{\pi}_i^t(j) \geq \gamma/|S_i| \geq \gamma/m$ for all $j \in S_i$ and all t, and form

$$\big[\widehat{\nabla}_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t)\big]_j = \begin{cases} u_i^t(\mathbf{s})/\tilde{\pi}_i^t(s_i), & j = s_i, \\ 0, & j \neq s_i. \end{cases} \tag{18}$$

This estimator is *sparse*: only the played action receives an observation, and the momentum in (6) accumulates evidence across rounds. Two consequences are carried through the analysis: the variance obeys $\tilde{\sigma}^2 \leq m^2 v_{\max}^2/\gamma$, and (18) is unbiased at $\tilde{\boldsymbol{\pi}}^t$. Corollary 2 therefore runs the whole argument at the played policy. Trading the two against each other fixes γ and the pure-bandit rate; under unilateral-oracle feedback no exploration is needed and Variant A runs with $\gamma = 0$ (Appendix A.6).

Lemma 7 (Appendix A.5) shows the projected momentum error of (6) obeys $\mathbb{E}\|e^t\|^2 \leq Q/(t+5)^{1/2}$ with $Q = 3n\tilde{\sigma}^2 + 10L^2D^2\eta^2T^{3/2} + 3nmv_{\max}^2$, which is free of T at $\eta = T^{-3/4}$.

A.4 Additional related work and table notes

Full notes to Table 1

[†]No concavity assumption is used (Proposition 1), and $\tilde{\sigma}^2, \sigma_H^2 = \mathrm{poly}(n,m)\,v_{\max}^2$ (Proposition 2). Per-round computation and memory are likewise $\mathrm{poly}(n,m)$ for general identity-dependent FVGs (Table 7); the $2^{\Theta(n)}$ cost of this class is *representational*. It is paid by any method that tabulates, regresses on, or identifies v_j, and is not paid by `OSD-FW` (Section B.5). [‡]Opt. B communicates no policies, statistics or private estimates, but does require *joint-action observability* of the realized allocation together with the two publicly announced auxiliary action draws per agent per round specified in Section 3.3.2; every such draw is generated by an agent from its own policy, so no agent ever needs another's distribution. Extension to pure bandit feedback is open (Section B.6). [¶]**Our pure-bandit row is superseded.** An FVG is a finite exact potential game with $|S_i| \leq m$ and payoffs in $[0, v_{\max}]$, so after dividing by $v_{\max}$ Dong et al. (2024) applies directly and attains $\mathcal{O}(T^{4/5})$, a better exponent than our $\mathcal{O}(T^{5/6})$, under the same pure-bandit feedback. Their bound is stated for general potential games and so carries no $v_{\max}$ or identity structure; the "Yes" in the last column means their result covers identity-dependent FVGs, not that it exploits their structure. We include their method as a baseline in Section 4. [‖]Conditional on the tracking hypothesis Assumption 2 for the windowed ridge estimator (21); the unconditional results of this paper are the three `OSD-FW` rows above. [§]Daskalakis et al. (2021) bound *external* regret, which certifies convergence of the empirical joint play to the set of coarse correlated equilibria; it is an external-regret guarantee, so the row locates the rate but is not a like-for-like comparison.

A.5 Additional technical details

Closing the momentum recursion

Two corrections relative to the standard argument. First, the recursion is closed with Young's inequality at parameter $\beta = \rho_t/2$, giving a leading coefficient $(1-\rho_t)$; the naive bound $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ leaves a coefficient $3(1-\rho_t)^2 > 1$ whenever $\rho_t < 2/3$ and yields an *expansive* recursion. Second, Q must be T-free for the error sum in Theorem 1 to decay; this is what pins η to $T^{-3/4}$ given $\rho_t \asymp t^{-1/2}$.

Variance and bias of the forced-exploration estimator

This estimator is *sparse*: only the played action s_i receives an observation per round; the momentum in (6) accumulates evidence across rounds for all components. Two consequences of (17) are carried explicitly through the analysis (Section 3.4):

- *Variance.* $\mathbb{E}\|\widehat{\nabla}_{\pi_i}\bar{\Phi}\|^2 \leq v_{\max}^2 \sum_{j\in S_i} 1/\tilde{\pi}_i^t(j) \leq m^2 v_{\max}^2/\gamma$, so $\tilde{\sigma}^2 \leq m^2 v_{\max}^2/\gamma$, bounded, but growing as $\gamma \to 0$.

- *Bias.* (18) is unbiased for $\nabla_{\pi_i} U_i(\tilde{\boldsymbol{\pi}}^t)$, not at $\boldsymbol{\pi}^t$, and the round's regret is incurred at $\tilde{\boldsymbol{\pi}}^t$. Since the total variation between $\tilde{\boldsymbol{\pi}}^t$ and $\boldsymbol{\pi}^t$ is at most $n\gamma$ and payoffs lie in $[0, v_{\max}]$, $\mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \le \mathsf{NG}(\boldsymbol{\pi}^t) + 4n\gamma v_{\max}$, an additive $4n\gamma v_{\max} T$ in Nash regret.

Trading these against each other is what fixes γ and, ultimately, the pure-bandit rate (Corollary 2). Under unilateral-oracle feedback no exploration is needed: (5) carries no importance weights, and Variant A is run with $\gamma = 0$.

PROPOSITION 1, PARTS (C) AND (D)

(c) A sufficient condition for non-concavity: if some pair $i \neq k$ shares two tasks $j, j' \in S_i \cap S_k$ with $H_{(i,j),(k,j)} + H_{(i,j'),(k,j')} \neq 0$, then $\bar{\Phi}$ is not concave on $\boldsymbol{\Delta}$. Under submodular v_j every such entry is ≤ 0, so the condition reduces to: two agents share two tasks and at least one of the two cross-partials is nonzero. This is the generic case.

(d) Under submodular v_j, $H_{(i,j),(k,j)} \le 0$ for $i \neq k$. Hence $\mathbf{w}^\top H \mathbf{w} \le 0$ for every $\mathbf{w} \ge 0$: $\bar{\Phi}$ is concave along non-negative directions (equivalently, continuous DR-submodular in the sense of Bian et al. (2017)). This, and not concavity, is the property submodularity buys.

Corollary 3. *To reach $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] \le \varepsilon$, Variant B requires*

$$T = \mathcal{O}\left(\frac{(mv_{\max} + n\tilde{\sigma} + n\sigma_H)^3}{\varepsilon^3}\right) = \mathcal{O}\left(\frac{n^6\, m^{3/2}\, v_{\max}^3}{\varepsilon^3}\right) \tag{19}$$

rounds under unilateral-oracle feedback with the auxiliary-draw protocol, using $\tilde{\sigma}^2 \le m v_{\max}^2$ and $\sigma_H^2 \le 4(n-1)^2 m v_{\max}^2$ (Proposition 2); the dominant term is $n\sigma_H = O(n^2\sqrt{m}\, v_{\max})$. The bound is log-free and polynomial in n and m, as are the per-round computation ($O(mn)$ oracle calls) and the per-agent memory ($O(m)$); see Table 7.

A.6 ADDITIONAL CONVERGENCE RESULTS

BANDIT ANALYSIS ON THE SHRUNKEN SIMPLEX

The point of routing the argument through $\boldsymbol{\Delta}^\gamma$ is that estimator (18) is conditionally unbiased for $\nabla_{\pi_i} U_i(\tilde{\boldsymbol{\pi}}^t)$, the gradient at the very policy the recursion tracks, so the momentum recursion of Lemma 7 is driven by mean-zero noise and carries no bias. Running the analysis at $\boldsymbol{\pi}^t$ instead and appending a Nash-gap comparison $\mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \le \mathsf{NG}(\boldsymbol{\pi}^t) + 4n\gamma v_{\max}$ would *not* suffice: it leaves unaccounted the bias sitting inside the momentum recursion and the ascent argument. Routed this way the exploration cost enters in one place only, the linear oracle ranges over the shrunken set $\boldsymbol{\Delta}^\gamma$, which loses at most $n\gamma v_{\max}$ of Frank–Wolfe gap per round relative to $\boldsymbol{\Delta}$ (Lemma 8).

Remark 7 (Why $T^{-1/4}$ and not $T^{-1/3}$)**.** ***We do not claim the exponent is a barrier.*** *It is* not *the case that an unweighted sum over rounds requires a constant step size. For $0 \le \bar{\Phi} \le B$ and any nonincreasing η_t, summation by parts gives $\sum_{t \le T}\big(\bar{\Phi}(\boldsymbol{\pi}^{t+1}) - \bar{\Phi}(\boldsymbol{\pi}^t)\big)/\eta_t \le B/\eta_T$, since the coefficients $1/\eta_{t-1} - 1/\eta_t$ are nonpositive and $\bar{\Phi} \ge 0$; the uniform average of $\mathsf{FW}(\boldsymbol{\pi}^t)$ can therefore be controlled with decreasing steps too. Dong et al. (2024) do exactly this, with $\eta_t \asymp t^{-2/5}$. What follows is thus a statement about the trade-off* within the constant-step family we analyse*, not about what is achievable. The telescope then contributes $mv_{\max}/(\eta T)$, pushing η* up*, while the momentum lag contributes $\asymp LD^2\eta T^p$ for $\rho_t \asymp t^{-p}$, pushing η* down*; the momentum noise contributes $\asymp D\tilde{\sigma}T^{-p/2}$. Optimizing over $(\eta, p) = (T^{-a}, p)$ gives $a = (1+p)/2$ and a rate exponent $\max\{(p-1)/2, -p/2\}$, minimized at $p = 1/2$, $a = 3/4$, value $-1/4$. The schedule $\eta = T^{-2/3}$, $\rho_t \asymp t^{-2/3}$ inherited from centralized SCG gives only $T^{-1/6}$ here, and an epoch-based estimator (freeze $\boldsymbol{\pi}$ for B rounds, average B samples) also gives $-1/4$. We report this as the boundary of the family we searched, and explicitly* not *as a lower bound: decreasing schedules are outside it, Table 11 shows a decreasing-style pairing beating ours empirically, and Dong et al. (2024) obtain $T^{-1/5}$ for the pure-bandit problem with decreasing steps. Whether a decreasing schedule improves the oracle-feedback exponents here is open.*

Corollary 4. *To reach $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)] \le \varepsilon$, Variant A under unilateral-oracle feedback requires*

$$T = \mathcal{O}\left(\frac{(n^2 m v_{\max} + n\tilde{\sigma})^4}{\varepsilon^4}\right) = \mathcal{O}\left(\frac{n^8 m^4 v_{\max}^4}{\varepsilon^4}\right) \tag{20}$$

rounds, using $\tilde{\sigma}^2 \le m v_{\max}^2$. This bound is log-free. Per-round computation and memory are polynomial in n and m as well, for general identity-dependent FVGs (Table 7, Section B.5); $O(m)$ oracle calls and $O(m)$ state per agent per round.

Lemma 8. *For every $\tilde{\boldsymbol{\pi}} \in \boldsymbol{\Delta}^\gamma$, $\mathsf{NG}(\tilde{\boldsymbol{\pi}}) \le \mathsf{FW}(\tilde{\boldsymbol{\pi}}) \le \mathsf{FW}^\gamma(\tilde{\boldsymbol{\pi}}) + n\gamma v_{\max}$, where $\mathsf{FW}^\gamma(\tilde{\boldsymbol{\pi}}) := \max_{\boldsymbol{\nu}\in\boldsymbol{\Delta}^\gamma} \langle \nabla\bar{\Phi}(\tilde{\boldsymbol{\pi}}), \boldsymbol{\nu} - \tilde{\boldsymbol{\pi}}\rangle$ is the Frank–Wolfe gap of the restricted oracle actually used.*

Proposition 2. *Each of the $n-1$ summands of (11) is a difference of two potentials at profiles differing in exactly one agent's action; such a move alters exactly two coalitions, so by Assumption 1 the summand lies in $[-2v_{\max}, 2v_{\max}]$. Hence $\left|[\tilde{\boldsymbol{\Delta}}_i^t]_j\right| \le 2\eta(n-1)v_{\max}$ for every $j \in S_i$, and since $|S_i| \le m$,*

$$\left\|\tilde{\boldsymbol{\Delta}}_i^t\right\| \;\le\; 2\eta(n-1)\sqrt{m}\, v_{\max}, \qquad \textit{so} \qquad \sigma_H^2 \;\le\; 4(n-1)^2 m\, v_{\max}^2.$$

The drift variance is therefore polynomial *in n and m for arbitrary identity-dependent coalition values; no property of v_j beyond boundedness is used, so the bound is insensitive to how identity-dependent the coalition values are. The same count bounds the per-round work: $|S_i|(n-1) \le m(n-1)$ potential differences, each two marginal contributions, i.e. $O(mn)$ value-oracle calls and $O(m)$ stored state.*

A.7 Additional experimental analysis

Throughout, the *gradient-path coherence* of a run is $\|\sum_t D_i^t\| / \sum_t \|D_i^t\|$ averaged over agents, where $D_i^t = \mathsf{P}_i(\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1}))$. Section 4 finds that the drift correction improves $\mathcal{R}^T/T$ in the low-coherence regime and not in the high-coherence one; Appendix C.1 says what a bound expressing that would have to look like, and why the obvious candidate has the sign backwards.

Settings behind Figure 1. Panel (a) uses $T \in \{250, 500, 1000, 2000, 4000\}$. Panel (b) pools 48 paired differences per environment: six configurations ($n \in \{4, 6, 8\}$ with $m \in \{4, 4, 5\}$, $T \in \{500, 2000\}$, noise 0.05) of eight paired runs each, the arms sharing instances, seeds and schedule and differing only in the drift term; each difference is taken relative to its own no-drift arm. Panel (c) uses n=8, m=5, T=2000 and eight runs per arm, with error bars of one standard error.

Gradient-path coherence: protocol and caveats. For a run we record $D_i^t = \mathsf{P}_i(\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1}))$, computed *exactly* by the mask summation of Appendix E, and report $\|\sum_t D_i^t\| / \sum_t \|D_i^t\|$ averaged over agents. The trajectory is the noiseless exact-gradient Frank–Wolfe path at the run's step size $\eta = T^{-2/3}$ with $T = 500$, so the statistic is a property of the game and the dynamics and not of any estimator. Measured over the same 36 instances the ablation used, the families separate without overlap — SKILL 0.82, COVERAGE 0.81, anonymous 0.07 — and the pooled association with the measured benefit is strong (Pearson $r = +0.76$, Spearman $\rho = +0.62$, permutation $p = 10^{-4}$, $n = 36$).

Three caveats keep us from calling coherence a predictor. First, that pooled association is driven entirely by the gap between the two clusters: restricted to the 24 identity-dependent instances, whose coherence still ranges over $[0.52, 1.00]$, it vanishes ($\rho = +0.14$, $p = 0.52$). Second, the anonymous generator is deterministic given (n, m), so its twelve points come from three value structures and their spread is run noise, not independent instances. Third, we have no argument that coherence is the operative variable rather than a correlate of one; that is the content of Open Problem 2. The measurement is reproducible via `drift_coherence.py` and `drift_predictor.py`.

All experiments below were produced by a single reference implementation under the corrected schedules of Algorithm 1; there are no legacy or exploratory results in the paper, and every number reported comes from the same run set.[3]

Protocol. Each configuration is run on 5 independently *resampled instances* of the environment (fresh skill matrix W, locations, or capability vectors) with 5 independent algorithm seeds each, giving 25 runs; the rate sweep uses 9. Reported $\pm$ figures are one standard deviation over all runs, so they capture instance variability as well as run-to-run variability. Unless stated, n=8, m=5, T=2000, and observation noise $\epsilon \sim \mathcal{N}(0, 0.05^2)$ clipped to $[0, v_{\max}]$.

[3] Code, raw result files, and plotting scripts are released with the submission.

Table 4: **Identity-dependent (skill-based) FVG**, n=8, m=5, T=2000, 5 instances $\times$ 5 seeds (25 runs). All quantities computed *exactly* on the policy played at every round. Rows are sorted by $\mathcal{R}^T/T = \frac{1}{T}\sum_t \mathsf{NG}(\tilde{\boldsymbol{\pi}}^t)$, which is the quantity Theorems 1–2 bound and equals $\mathbb{E}[\mathsf{NG}(\tilde{\boldsymbol{\pi}}^\tau)]$; cumulative regret is T times it. "Calls/rd" is value-oracle calls per round summed over agents. Bold = best per column. $\star$ marks methods with access others do not have.

Algorithm	Feedback	$\mathcal{R}^T/T \downarrow$	Final gap $\downarrow$	Cumul. regret $\downarrow$	Social welfare $\uparrow$	Calls/rd
MF-OMD*	Full-info	$\mathbf{0.0150 \pm 0.0033}$	0.0006 ± 0.0006	$\mathbf{30.0 \pm 6.6}$	4.317 ± 0.203	0
IndepPG	Unilateral-oracle	0.0190 ± 0.0052	0.0056 ± 0.0058	38.0 ± 10.5	$\mathbf{4.369 \pm 0.183}$	80
ε-Greedy	Unilateral-oracle	0.0416 ± 0.0043	0.0168 ± 0.0100	83.2 ± 8.5	4.281 ± 0.204	8
OSD−FW-B	Aux.-draw oracle	0.0416 ± 0.0024	$\mathbf{0.0001 \pm 0.0002}$	83.3 ± 4.9	4.298 ± 0.216	165
UCB-MC	Unilateral-oracle	0.0541 ± 0.0061	0.0142 ± 0.0325	108.3 ± 12.2	4.332 ± 0.237	8
OSD−FW-A	Unilateral-oracle	0.0770 ± 0.0033	0.0015 ± 0.0010	154.0 ± 6.6	4.329 ± 0.195	80
IndepMD	Unilateral-oracle	0.0873 ± 0.0032	0.0331 ± 0.0051	174.6 ± 6.4	4.317 ± 0.166	80
FW-Exploration	Unilateral-oracle	0.1009 ± 0.0040	0.0928 ± 0.0038	201.8 ± 8.0	4.005 ± 0.131	80
OSD−FW-A (γ=0.1)	Bandit	0.1788 ± 0.0118	0.1137 ± 0.0177	357.5 ± 23.6	3.855 ± 0.139	8
OSD−FW-A ($\gamma^\star$)	Bandit	0.2313 ± 0.0169	0.1878 ± 0.0162	462.7 ± 33.8	3.458 ± 0.089	8

Table 5: **Anonymous FVG**, same protocol and layout as Table 4.

Algorithm	Feedback	$\mathcal{R}^T/T \downarrow$	Final gap $\downarrow$	Cumul. regret $\downarrow$	Social welfare $\uparrow$	Calls/rd
MF-OMD*	Full-info	$\mathbf{0.0046 \pm 0.0000}$	$\mathbf{0.0000 \pm 0.0000}$	$\mathbf{9.1 \pm 0.0}$	3.230 ± 0.000	0
ε-Greedy	Unilateral-oracle	0.0182 ± 0.0014	0.0098 ± 0.0000	36.5 ± 2.8	3.673 ± 0.000	8
OSD−FW-B	Aux.-draw oracle	0.0243 ± 0.0015	0.0100 ± 0.0056	48.6 ± 3.1	3.792 ± 0.076	232
IndepMD	Unilateral-oracle	0.0255 ± 0.0028	0.0107 ± 0.0032	51.1 ± 5.6	3.270 ± 0.003	80
IndepPG	Unilateral-oracle	0.0431 ± 0.0043	0.0406 ± 0.0107	86.1 ± 8.5	3.508 ± 0.107	80
OSD−FW-A	Unilateral-oracle	0.0433 ± 0.0018	0.0308 ± 0.0101	86.5 ± 3.7	3.631 ± 0.104	80
FW-Exploration	Unilateral-oracle	0.0444 ± 0.0014	0.0386 ± 0.0106	88.9 ± 2.9	3.303 ± 0.022	80
UCB-MC	Unilateral-oracle	0.0884 ± 0.0045	0.0348 ± 0.0362	176.8 ± 8.9	$\mathbf{3.833 \pm 0.211}$	8
OSD−FW-A (γ=0.1)	Bandit	0.1051 ± 0.0031	0.0777 ± 0.0129	210.1 ± 6.3	3.360 ± 0.004	8
OSD−FW-A ($\gamma^\star$)	Bandit	0.1205 ± 0.0024	0.0917 ± 0.0104	241.0 ± 4.8	3.360 ± 0.002	8

Evaluation metrics. Nash gap, social welfare and potential are computed *exactly*, by summation over coalition masks, using the identity $\mathbb{E}_{\mathbf{s}_{-i}}[u_i(j, \mathbf{s}_{-i})] = \sum_{C \subseteq N\setminus\{i\}} \mathbb{P}(C_j(\mathbf{s}_{-i}) = C)\,[\bar{v}_j(C \cup \{i\}) - \bar{v}_j(C)]$, at cost $\Theta(nm2^n)$ per evaluation, and they are computed *every round* on the policy actually *played*. The quantity bounded by the theorems is

$$\mathcal{R}^T/T \;=\; \tfrac{1}{T}\sum_{t\le T} \mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \;=\; \mathbb{E}\big[\mathsf{NG}(\tilde{\boldsymbol{\pi}}^\tau)\big], \qquad \tau \sim \mathrm{Unif}\{1,\dots,T\},$$

This differs from the Nash gap of the final or best iterate, a distinction that changes the conclusions materially (see the empirical-rate analysis below). We therefore report $\mathcal{R}^T/T$ as the primary column throughout; cumulative regret is exactly T times it. We also report the final-iterate gap, because it answers a different and legitimate question (where does the method end up?), and the per-round value-oracle budget, because the methods compared do not spend the same one.

Algorithms. Ten methods: (i)–(iv) OSD−FW-A (unilateral-oracle), OSD−FW-A (bandit) at $\gamma^\star = (m/4)^{2/3}T^{-1/6}$, one member of the $\Theta(m^{2/3}T^{-1/6})$ family of Corollary 2, and at a tuned γ=0.1, and OSD−FW-B (auxiliary-draw oracle, with the sampled drift correction (11)); (v) *UCB-MC*, a decentralized UCB baseline in which each agent keeps per-(agent, task) counts and means of her realized marginal contribution and greedily plays the largest upper confidence bound (α=0.3), the natural FVG analogue of the UCB-based equilibrium learners of Cui et al. (2022); Cohen & Agmon (2024), but neither of those algorithms and not centralized; (vi) *FW-Exploration*, mini-batch Frank–Wolfe with forced exploration in the style of Cui et al. (2022); (vii) *ε-Greedy*; (viii) *IndepPG*, independent projected gradient ascent; (ix) *IndepMD*, independent exponentiated-gradient mirror descent; (x) *MF-OMD*, optimistic mirror descent supplied with the *exact* expected marginal contributions from the evaluator, included as an oracle reference point rather than as a competitor: it reads information no other method here has, and is not a decentralized learner; (xi) *Dong et al. (2024), Algorithm 1*, the closest prior method, a faithful transcription of their potential-game bandit algorithm at their Theorem 3.1 parameters ($\eta_t = \sqrt{1/(nt^{4/5})}$, $\mu = \min\{1/(mn), m/(n^{1/8}T^{1/5})\}$, $\rho_t = 4\mu^{1/3}n^{1/3}m^{1/3}t^{-8/15}$), with payoffs normalised to $[0,1]$ by dividing by $v_{\max}$, which is what makes an FVG an instance of their class. We clip ρ_t at 1; they state no clip, and at these constants ρ_t

exceeds 1 for small t. All results below were produced with the revised implementation, in which Variant B uses the sampled drift correction. The previous implementation formed the Jacobian–vector product against $\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}$ directly and so was not decentralized; its numbers do not carry over and are not reused anywhere in this paper. *D1S-FW* (Cohen & Agmon, 2025a) is not evaluated: its G-optimal design is built on the pairwise-linear parameterization of additively separable hedonic games, which FVG coalition values do not admit, so there is no faithful adaptation to run. It appears in Table 1 as a theoretical comparison only.

Throughout the tables and figure legends we use the feedback names fixed in Section 3.1 and avoid the word "semi-bandit", which in the bandit literature means observing the components of one's own played action and is weaker than what these methods use. "Unilateral-oracle" means joint-action observability of the realized allocation plus a value oracle on single-deviation coalitions; "Aux.-draw oracle" means that plus the auxiliary announcements of Section 3.3.2 and oracle access at the assembled profiles, which are not confined to a neighbourhood of the played allocation. Only `OSD-FW`-A under "bandit" feedback uses own-payoff information alone. The "Calls/rd" column makes the resulting budget difference explicit.

Environments. All environments are synthetic and fully specified; none uses field data. *Anonymous*: $v_j(C) = a_j\sqrt{|C|}$, $a_j \sim \mathrm{Unif}[0.4, 1.5]$. *Skill-based (identity-dependent)*: $v_j(C) = \sqrt{\sum_{i\in C} W_{ij}}$ with $W_{ij} \sim \mathrm{Exp}(1)$ normalised so that $\max_i W_{ij} = 1$ for each task, a concave increasing function of a modular set function, hence monotone *and* submodular, so this instance satisfies every assumption in the paper while being genuinely identity-dependent. *Crowdsourcing-style* and *MRTA-style*: as above but with W_{ij} given by spatial proximity and by capability–requirement coverage respectively, and with heterogeneous feasible sets S_i induced by the same quantities, in the spirit of the settings of Chen et al. (2014) and Bichler et al. (2025); Yan & Di (2022). *Non-monotone stress test*: $v_j(C) = (\sum_{i\in C} W_{ij})/\sqrt{|C|}$, which is *not* monotone, adding an agent with $W_{ij} < (\sum_{k\in C} W_{kj})(\sqrt{1 + 1/|C|} - 1)$ lowers the value, and therefore lies outside Definition 1's hypotheses; see the non-monotone stress test in Appendix E.

Interpretation of results. The average and final-iterate gaps rank the methods differently, so we report both. On the identity-dependent instance at equal T, `OSD-FW`-B attains the lowest *final-iterate* Nash gap by a wide margin, 0.0001 ± 0.0002, against 0.0015 ± 0.0010 for `OSD-FW`-A and below every baseline including the exact-gradient reference MF-OMD (0.0006 ± 0.0006; Table 4). The controlled ablation of Section 4 shows that this equal-T gap between the two variants is a step-size effect rather than the signature of the drift correction, which is why we do not read it as evidence for the estimator. On $\mathcal{R}^T/T$, the primary objective, `OSD-FW`-B is *fourth* at 0.0416 ± 0.0024: behind MF-OMD (0.0150) and IndepPG (0.0190 ± 0.0052), and tied with ε-Greedy (0.0416 ± 0.0043). Frank–Wolfe with a constant step moves slowly toward a vertex, so many early rounds sit far from equilibrium and contribute to $\mathcal{R}^T/T$ even when the final gap is small; methods with a larger early step show the opposite pattern, IndepPG accumulating less than half the regret and ending at a final gap roughly $50\times$ larger. Which is preferable depends on whether the deployment cares about cost incurred while learning or about the profile it ends up running. Against the closest prior method, Dong et al. (2024) reach $\mathcal{R}^T/T = 0.2147$ at this horizon, behind both of our bandit configurations (0.1788 and 0.2313). That ordering does *not* survive longer horizons or a budget match and should not be read as an advantage: their exponent is better (Figure 1a), and at the matched budget of Table 6 they finish at 0.0860 against our bandit variant's 0.1138. This comparison also does not equalize oracle budgets. `OSD-FW`-B spends 165 value-oracle calls per round summed over agents, against 80 for `OSD-FW`-A, IndepPG, IndepMD and FW-Exploration, and 8 for the pure-bandit, UCB-MC and ε-Greedy methods; MF-OMD spends none, being handed exact gradients. Per call, B's final-iterate advantage is smaller than the raw table suggests and its $\mathcal{R}^T/T$ disadvantage larger — which is what Table 6 corrects for.

Matched oracle budget. Table 6 fixes a total budget of $\approx 1.98 \times 10^5$ value-oracle calls and gives every method the horizon its own per-round rate affords, so the scarce resource is charged equally instead of the round count being equalised; cheap methods consequently run 20–22$\times$ longer. Section 4 reads off the ranking. Two details belong here: UCB-MC locks onto a pure profile over 32,933 rounds and uses own-payoff feedback only, so no final-iterate claim restricted to "decentralized methods" is available to us; and Dong et al. (2024) place ninth of ten on $\mathcal{R}^T/T$ (0.0860), ahead of our own pure-bandit variant (0.1138), consistent with their better exponent.

Table 6: **Matched value-oracle budget, enforced exactly.** Identity-dependent FVG, n=6, m=4, 9 runs. Each method runs until it has spent the same number of value-oracle calls (spread 11.8%); because the step size depends on T, the horizon is obtained from the fixed point $T \leftarrow T \cdot \text{budget}/\text{spend}(T)$ rather than from a probe. Metrics are exact over each method's own horizon; $\pm$ is one standard error. **UCB-MC attains the lowest mean on *both* columns.** IndepMD (0.0400) and FW-Exploration (0.0772) are in Appendix A.7.

Method	Feedback	T	Calls	$\mathcal{R}^T/T \downarrow$	Final gap $\downarrow$
UCB-MC	Bandit	32,933	197,598	$\mathbf{0.0047 \pm 0.0005}$	$\mathbf{0.00000 \pm 0.00000}$
IndepPG	Unilateral	4,117	197,616	0.0075 ± 0.0004	0.00102 ± 0.00063
ε-Greedy	Bandit	32,933	197,598	0.0130 ± 0.0004	0.01171 ± 0.00037
OSD−FW-B (sampled)	Aux.-draw	2,060	174,288	0.0342 ± 0.0011	0.00033 ± 0.00020
OSD−FW-A	Unilateral	4,117	197,616	0.0530 ± 0.0006	0.00099 ± 0.00043
OSD−FW-B (exact JVP)	Aux.-draw	375	197,520	0.0626 ± 0.0025	0.00145 ± 0.00058
Dong et al. (2024)	Bandit	32,933	197,598	0.0860 ± 0.0030	0.04532 ± 0.00588
OSD−FW-A (bandit)	Bandit	32,933	197,598	0.1138 ± 0.0024	0.08859 ± 0.00173

On the anonymous instance (Table 5), OSD−FW-B's advantage largely disappears. It is second on $\mathcal{R}^T/T$ among the decentralized methods (0.0243 ± 0.0015, behind ε-Greedy at 0.0182 ± 0.0014) and its final gap, 0.0100 ± 0.0056, is statistically indistinguishable from ε-Greedy's 0.0098 and IndepMD's 0.0107, a very different picture from the identity-dependent instance, where it led by an order of magnitude. Note that this is a different comparison from the ablation of Section 4, and the two are consistent: *within* Variant B the drift term is worth 16.3% on anonymous instances, while *across* methods Variant B is not the best choice there, because a method exploiting coalition-size structure directly has no handicap. Both readings point to the anonymous specialization of Section A.8 rather than to general Variant B on such instances. Where OSD−FW-B does stay ahead is social welfare (3.792, second only to UCB-MC's 3.833 and well above ε-Greedy's 3.673), and note more generally that low regret does not imply high welfare: MF-OMD accumulates by far the least regret while converging to the lowest-welfare equilibrium in the table (3.230).

Empirical rates. Figure 1a regresses $\log(\mathcal{R}^T/T)$ on $\log T$ over $T \in \{250, \ldots, 4000\}$, n=6, m=4, 9 runs per point. All four methods track their certified exponents closely:

	A (unilateral)	B	A (bandit)	Dong et al. (2024)
Fitted slope	-0.239	-0.335	-0.151	-0.211
Certified exponent	$-1/4$	$-1/3$	$-1/6$	$-1/5$
R^2	1.000	1.000	0.995	1.000

Fitting the *final-iterate* gap instead gives slopes of roughly -1.4 and -1.0 for the two oracle-feedback variants, which would suggest the bounds are loose; that reading is an artifact of fitting a quantity the theorems do not bound. On the same runs, the final-iterate slopes are -1.41, -0.98 and -0.19, while the average-gap slopes are the ones displayed above; the theorems are about the second row, and on it the fitted and certified exponents agree to within 0.016. Fitted slopes over a sixteen-fold range of T show the bounds are *consistent with* the observed decay, not that the exponents are tight; that would need a matching lower bound (Remark 7). What the final-iterate slopes do show is a separate and unclaimed phenomenon: once the Frank–Wolfe direction stabilizes, convergence of the last iterate to the equilibrium vertex is far faster than the average-gap rate, which is consistent with the certified exponents being driven by the transient. A last-iterate analysis is not attempted here.

Beyond the assumptions. On the non-monotone environment of Appendix E, where Definition 1's hypotheses fail, both OSD−FW variants still converge (final gaps 0.0050 ± 0.0140 and 0.0033 ± 0.0028) while FW-Exploration stalls at 0.0923 ± 0.0079. This is consistent with Remark 2: Lemma 1 and Corollary 1 hold for arbitrary v_j, and only the explicit smoothness constant uses submodularity.

Ablations on the exploration weight, momentum-error decay, scalability in n, the feasibility-structured environments, and a corrected-versus-original schedule comparison are in Appendix E.

A.8 ANONYMOUS FVGS: A COMPACT LINEAR PARAMETERISATION

OSD−FW already runs in $\mathrm{poly}(n, m)$ time and memory per round on general identity-dependent FVGs (Proposition 2, Table 7), so the purpose of this section is *not* to remove an exponential per-round cost, there is none to remove. What anonymity buys is a compact *representation*: the $2^{\Theta(n)}$-dimensional hypothesis class collapses to $m(n+1)$ parameters. That makes a model-based, regression-driven variant of the gradient and drift estimators available, which is unavailable in the general case, and it tightens the $\mathrm{poly}(n, m)$ constants in the regret bounds. The exponents in T are unchanged.

Definition 2. *An FVG is* anonymous *if $v_j(C) = f_j(|C|)$ for some $f_j : \{0, \dots, n\} \to [0, v_{\max}]$ with $f_j(0) = 0$, for every task j.*

Under anonymity, the marginal contribution $\delta_j(k) := f_j(k) - f_j(k-1)$ depends only on the coalition size k. This parallels congestion games Rosenthal (1973); Cui et al. (2022).

Linear parameterization. For agent i and task j, define $\phi_i(j) \in \mathbb{R}^{m(n+1)}$ encoding task j and the load distribution at j under $\boldsymbol{\pi}_{-i}$. Then

$$\mathbb{E}_{\mathbf{s}_{-i}}[u_i(j, \mathbf{s}_{-i})] = \sum_{k=0}^{n-1} \delta_j(k+1)\, \mathbb{P}_{\mathbf{s}_{-i}}(|C_j(\mathbf{s}_{-i})| = k) = \langle \phi_i(j), \theta(\boldsymbol{\pi}_{-i}) \rangle,$$

where $\theta(\boldsymbol{\pi}_{-i}) \in \mathbb{R}^{m(n+1)}$ collects marginal contributions weighted by load probabilities. By Lemma 2, $\nabla_{\pi_i} \bar{\Phi}(\boldsymbol{\pi})$ is linear in ϕ_i up to an additive constant absorbed by the FW oracle. There are only $m(n+1)$ parameters, so unlike the general case the gradient is *representable* by a model an agent can fit This finite representation is what the estimator below exploits. The claim concerns representation, not about OSD−FW's per-round cost, which is already polynomial in both cases (Section B.5).

Ridge-regression gradient estimator. Fix agent i and write $d = m(n+1)$. At round r she observes the realized loads $(|C_j(\mathbf{s}^r)|)_{j \in M}$ and her own realized marginal contribution $y^r = u_i^r(\mathbf{s}^r)$, and forms the feature $\phi^r := \phi_i(s_i^r) \in \mathbb{R}^d$ described above. Over a trailing window of length W she computes the regularized least-squares estimate

$$\hat{\theta}^t = \Big(\lambda I + \sum_{r=t-W+1}^{t} \phi^r \phi^{r\top}\Big)^{-1} \sum_{r=t-W+1}^{t} \phi^r y^r, \tag{21}$$

with the played policy mixed, as in (17), against the G-optimal design on $\{\phi_i(j)\}_{j \in S_i}$, which is what keeps $\lambda_{\min}$ of the design matrix growing linearly in W.

Non-stationary regression target. The regression target $\theta(\boldsymbol{\pi}_{-i}^t)$ is *not* fixed: it moves as the opponents' policies move. By (8) the joint policy travels at most ηD per round, so $\|\theta(\boldsymbol{\pi}_{-i}^t) - \theta(\boldsymbol{\pi}_{-i}^{t-1})\| = O(\eta\, \mathrm{poly}(n, m)\, v_{\max})$; a windowed estimator therefore carries a *tracking bias* of order $W\eta\, \mathrm{poly}(n, m) v_{\max}$ alongside its $O(d\, v_{\max}^2 / W)$ variance, and balancing the two suggests $W = \Theta(\eta^{-2/3})$. The familiar $\mathcal{O}(\mathrm{poly}(n, m) v_{\max}^2 / t)$ rate is the *stationary*-target rate and does not apply as stated. We do not have a complete analysis of (21) coupled to the Frank–Wolfe dynamics, so we state what the specialization needs as an explicit hypothesis and mark the result conditional.

Assumption 2 (Tracking, for the anonymous specialization only)**.** *The windowed estimator (21) run alongside Algorithm 1 satisfies $\mathbb{E}\|\hat{\theta}^t - \theta(\boldsymbol{\pi}_{-i}^t)\|^2 \le C\, \mathrm{poly}(n, m)\, v_{\max}^2\, \varrho_t$ for a deterministic sequence $\varrho_t \to 0$ with $\frac{1}{T}\sum_{t \le T} \sqrt{\varrho_t} = O(T^{-1/4})$ for Variant A and $O(T^{-1/3})$ for Variant B.*

Drift correction without auxiliary draws. Under the linear structure $\nabla_{\pi_i} \bar{\Phi}(\boldsymbol{\pi})(j) = \langle \phi_i(j), \theta(\boldsymbol{\pi}_{-i}) \rangle + \text{const}$, the drift can be estimated directly by $\hat{\theta}^t - \hat{\theta}^{t-1}$, with no auxiliary announcements and no sampled potential differences, giving $\sigma_H^2 = \mathcal{O}(\mathrm{poly}(n, m)\, v_{\max}^2)$ under Assumption 2. Note this makes the anonymous specialization *more* decentralized than general Variant B, not less: it needs the realized loads only.

Corollary 5 (conditional on Assumption 2)**.** *For anonymous FVGs with the ridge-regression estimator:*

(a) Variant A (unilateral-oracle):

$$T = \mathcal{O}\left(\frac{\mathrm{poly}(n,m)\, v_{\max}^4}{\varepsilon^4}\right) \tag{22}$$

(b) Variant B (unilateral-oracle):

$$T = \mathcal{O}\left(\frac{\mathrm{poly}(n,m)\, v_{\max}^3}{\varepsilon^3}\right) \tag{23}$$

Both bounds are polynomial in n and m. Per-round computation is polynomial in the general case too (Table 7); what is specific to anonymity is that the represented *model has $m(n+1)$ entries, compared with up to 2^n per task in the general case. This is what makes (21) possible and what improves the $\mathrm{poly}(n,m)$ constants. This corollary is conditional on Assumption 2. Proving that assumption for (21) under Frank–Wolfe dynamics is left open, and the corresponding rows of Table 1 are marked accordingly.*

Polynomial factors in sample complexity. Nash regret measures vanishing exploitability, while sample complexity expresses the same guarantee in terms of a round budget. When each round requires a physical deployment, a paid batch of tasks, or an A/B epoch, feasibility depends on both the rate exponent and its polynomial factors. In (22)–(23), the dominant factor is the G-optimal-design dimension $d = m(n+1)$ of the anonymous linear parameterization. For a stationary target, the ridge estimator would satisfy $\mathbb{E}\|\hat{\theta}^t - \theta\|^2 = \mathcal{O}(d\, v_{\max}^2/t) = \mathcal{O}(mn\, v_{\max}^2/t)$. Here the target changes over time, so these rows depend on Assumption 2 and are marked conditional. The anonymous specialization has per-round design cost $\mathcal{O}(d^2) = \mathcal{O}(m^2n^2)$. Table 9 measures a different cost: it runs Variant A under unilateral-oracle feedback, with no fitted model or G-optimal design, and exact evaluation dominates its running time at $\Theta(nm2^n)$ (Appendix E). These explicit exponents permit direct comparison with the n, F dependence of the congestion-game baselines in Table 1.

Table 7: Per-agent resource cost. *Every row is polynomial in n and m in both columns*: for `OSD-FW` there is no computational separation between general and anonymous FVGs, and we no longer claim one. The last row is the size of the model a *model-based* method would have to represent or identify, a cost `OSD-FW` does not pay, since it consumes each oracle response once and caches nothing.

Quantity	**General FVG**	**Anonymous FVG**
Momentum state $\mathbf{d}_i^t$	$O(\lvert S_i\rvert) \le O(m)$	$O(m)$
Total memory, Variant B	$O(m+n)$	$O(m+n)$
Samples/round, Variant A (bandit)	$O(1)$	$O(1)$
Oracle queries/round, Variant A (unilateral)	$O(\lvert S_i\rvert) \le O(m)$	$O(m)$
Oracle queries/round, Variant B	$O(mn)$	$O(mn)$
Arithmetic/round, Variant B	$O(mn)$	$O(m^2n^2)^*$
Gradient variance $\tilde{\sigma}^2$	$\le mv_{\max}^2$	$\le mv_{\max}^2$
Drift variance σ_H^2	$\le 4n^2mv_{\max}^2$	$O(\mathrm{poly}(n,m)v_{\max}^2)$
Represented model size (model-based methods only)	$\le m2^n$	$m(n+1)$

*The anonymous column is *larger* here because the G-optimal design solves a $d \times d$ system with $d = m(n+1)$. It buys smaller constants and a representable model, not a cheaper round. This cost is not what Table 9 measures; that figure runs Variant A, which fits no model.

B Supporting remarks

B.1 Potential function range

The potential $\Phi(\mathbf{s}) = \sum_j v_j(C_j(\mathbf{s}))$ is the total task value, *not* the social welfare $\sum_i u_i(\mathbf{s})$. The two coincide only for additive value functions; in general, an agent's marginal contribution is strictly less

than the coalition value. Consequently the range of $\bar{\Phi}$ on $\mathbf{\Delta}$ satisfies $0 \le \bar{\Phi}(\boldsymbol{\pi}) \le \sum_j \bar{v}_j(N_j) \le mv_{\max}$. All telescoping arguments in the convergence proofs (Steps 4 of Theorems 1 and 2) use this bound; using $nv_{\max}$ in its place would be incorrect.

B.2 Gradient estimation and the constant shift

Lemma 2 shows that $\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}) = \nabla_{\pi_i}U_i(\boldsymbol{\pi}) + c_i(\boldsymbol{\pi}_{-i})\mathbf{1}$ where $|c_i| \le (m+1)v_{\max}$, and the estimators (5) and (18) target $\nabla_{\pi_i}U_i$. The potential gradient differs by the constant shift $c_i(\boldsymbol{\pi}_{-i})\mathbf{1}$, which cancels exactly. An alternative accounting would inflate every downstream constant by an $\mathcal{O}(m^3 v_{\max}^2)$ term that does not belong there.

Let $\mathsf{P}_i = I - \frac{1}{|S_i|}\mathbf{1}\mathbf{1}^\top$. Every place a gradient or gradient estimate enters the analysis, it enters through an inner product with a difference of two points of $\Delta(S_i)$:

- the linear oracle, $\arg\max_{\mu_i}\langle \mathbf{d}_i^t, \mu_i\rangle$, whose solution set depends on $\mathbf{d}_i^t$ only through $\mathsf{P}_i\mathbf{d}_i^t$;
- the Frank–Wolfe gap (4), a maximum of $\langle \nabla_{\pi_i}\bar{\Phi}, \mu_i - \pi_i\rangle$;
- Step 2 of the proof of Theorem 1, where the error term is $\langle \nabla_{\pi_i}\bar{\Phi} - \mathbf{d}_i^t,\ \mu_i^t - \pi_i^t\rangle$.

In each case $\langle g, \mu_i - \pi_i\rangle = \langle \mathsf{P}_i g, \mu_i - \pi_i\rangle$, because $\mu_i - \pi_i \perp \mathbf{1}$. Defining the error as the *projected* error $e_i^t := \mathsf{P}_i(\mathbf{d}_i^t - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t))$ therefore loses nothing, and since $\mathsf{P}_i(c_i\mathbf{1}) = \mathbf{0}$,

$$e_i^t \;=\; \mathsf{P}_i\big(\mathbf{d}_i^t - \nabla_{\pi_i}U_i(\boldsymbol{\pi}^t)\big), \qquad \mathbb{E}\big\|\mathsf{P}_i(\widehat{\nabla}_{\pi_i}\bar{\Phi} - \nabla_{\pi_i}\bar{\Phi})\big\|^2 = \mathbb{E}\big\|\mathsf{P}_i(\widehat{\nabla}_{\pi_i}U_i - \nabla_{\pi_i}U_i)\big\|^2 \le \tilde{\sigma}^2.$$

The momentum recursion is linear, so P_i commutes with it and the recursion for e_i^t is driven by the mean-zero noise $\mathsf{P}_i(\widehat{\nabla}_{\pi_i}U_i - \nabla_{\pi_i}U_i)$ alone. The effective variance is $\tilde{\sigma}^2$; there is no additive $|S_i|c_i^2$ penalty anywhere, and the constants Q, Q_B of Lemma 7 and Theorem 2 contain no $m^3 v_{\max}^2$ term.

B.3 Sampling protocol for Variants A and B

The estimators (5)–(11) evaluate the value oracle at several coalitions per round, and the mean-square recursions of Lemma 7 and Theorem 2 depend on which realizations are shared and which are independent. We fix the protocol explicitly.

Round t.

(S1) *Announcements (Variant B only).* Before play, each agent k publishes b_k^t, the task supported by her own Frank–Wolfe vertex μ_k^{t-1}, and a fresh draw $a_k^t \sim \pi_k^{t-1}$ generated by agent k's own randomizer, independently across agents and independently of everything else given $\mathcal{F}_{t-1}$. Each is a single action; neither is a distribution.

(S2) *Play draw.* Each agent samples $s_i^t \sim \tilde{\pi}_i^t$ once, independently across agents. This single profile is used both to collect the round's payoff and to index the gradient estimator; no second profile is played.

(S3) *Drift draw (Variant B only).* Agent i draws $\alpha^t \sim \mathrm{Unif}[0,1]$ and coins $c_k^t \sim \mathrm{Bern}(\alpha^t\eta)$ from her *own* randomizer, independently of (S2) and of $\mathcal{F}_{t-1}$, and assembles $\mathbf{s}'^t$ from the announcements by (10). $\mathbf{s}'^t$ is not played; it indexes counterfactual oracle queries only. Note what agent i does and does not sample: she samples her own α^t and coins, never another agent's policy. The randomness that carries π_k^{t-1} and μ_k^{t-1} into the estimator was generated by agent k in (S1).

(S4) *Oracle noise.* Every oracle *call* returns an independent noise realization: $v_j^t(C) = \bar{v}_j(C) + \epsilon_{j,C,\iota}^t$ with $\{\epsilon_{j,C,\iota}^t\}$ independent across calls ι, across pairs (j,C), and across rounds, and independent of (S1)–(S3). In particular the gradient-query stream and the drift-query stream receive independent noise *even where they touch the same pair* (j,C).

Role of each sampling choice. Independence in (S4) makes the estimators unbiased for the mean coalition values and gives the variance bounds $\tilde{\sigma}^2 \le mv_{\max}^2$ (unilateral-oracle) and $\tilde{\sigma}^2 \le m^2 v_{\max}^2/\gamma$ (bandit) used throughout. The independence of (S3) from (S2) kills the term $\mathbb{E}\langle e^{t-1}, \mathsf{P}\xi^t\rangle$ in Step 3 of the proof of Theorem 2.

Shared oracle noise: what independence of the draws does *not* buy. Reusing one noise realization per distinct (j, C) across *both* estimators, and appealing to independence of the profile draws, does *not* kill the whole cross term $\mathbb{E}\langle e^{t-1} + \zeta^t, \mathsf{P}\xi^t\rangle$: independent profile draws leave ζ^t and ξ^t coupled through the shared noise variable, so $\mathbb{E}\langle\zeta^t, \mathsf{P}\xi^t\rangle$ need not vanish. We resolve this in two ways, and record both. *(i) By design.* (S4) as now stated draws independent noise per call, so $\mathsf{P}\xi^t$ is conditionally mean-zero given $\sigma(\mathcal{F}_{t-1}, \alpha^t, \{c_k^t\}, \{a_k^t, b_k^t\}, \text{drift-query noise})$, and the cross term vanishes exactly. This costs nothing, each call is a separate evaluation of the scoring model, and it is what we implement. *(ii) Without it.* If common random numbers are used, Cauchy–Schwarz gives $|\mathbb{E}\langle\zeta^t, \mathsf{P}\xi^t\rangle| \le \sqrt{\mathbb{E}\|\zeta^t\|^2}\sqrt{\mathbb{E}\|\mathsf{P}\xi^t\|^2} \le n\eta\sigma_H\tilde{\sigma}$, so (31) acquires an additive $2\rho_t n\eta\sigma_H\tilde{\sigma}$. Since $\rho_t\eta = (t+1)^{-2/3}T^{-2/3} \le (t+1)^{-4/3}$ for every $t \le T$, this term sits at the $(t+1)^{-4/3}$ scale the induction of Step 4 already runs at, and is absorbed by enlarging Q_B with an additive $6n\sigma_H\tilde{\sigma} \le 3n(\sigma_H^2 + \tilde{\sigma}^2)$. The exponent is unchanged and the constants degrade by a constant factor. The guarantee therefore does not *depend* on (S4)(i); we prefer it because it is free.

Query accounting per round. Variant A pure bandit: one payoff, no queries. Variant A unilateral-oracle: $|S_i| \le m$ marginal contributions, i.e. $\le 2m$ oracle calls, on single-deviation coalitions. Variant B also computes $|S_i|(n-1) \le m(n-1)$ potential differences. Each such difference is taken at two profiles agreeing outside agent k, so by $\Phi(j, \ell, \mathbf{s}_{-\{i,k\}}) - \Phi(j, \ell', \mathbf{s}_{-\{i,k\}}) = \delta_{k,\ell}(B_\ell \cup \{k\}) - \delta_{k,\ell'}(B_{\ell'} \cup \{k\})$ it costs two marginal contributions, giving $O(mn)$ oracle calls in total, a factor m cheaper than the Jacobian–vector form it replaces, because the baseline column that the old estimator summed over ℓ cancels identically against $\sum_\ell(\pi_k^t(\ell) - \pi_k^{t-1}(\ell)) = 0$.

Where the query profiles live. The coalitions queried by Variant A lie in the unilateral-deviation neighbourhood of the played profile. The coalitions queried by Variant B do *not*: $\mathbf{s}'^t$ is assembled from announcements and may differ from $\mathbf{s}^t$ in many coordinates, so Variant B does *not* query only coalitions within a bounded Hamming distance of the realized profile. What is true, and what the algorithm needs, is that (i) the *number* of queries is $O(mn)$, (ii) every queried profile is assembled from publicly announced single actions, and (iii) the oracle is a value oracle on sets, i.e. the platform's scoring model $\hat{v}_j(\cdot)$, which by assumption can be evaluated at any candidate team (Section 3.1). Variant B is appropriate exactly when such a model exists, and not when the only way to learn $v_j(C)$ is to physically form C.

B.4 Concavity of $\bar{\Phi}$

Proof of Proposition 1. *(a)* Extend $\bar{\Phi}$ multilinearly to the ambient box, so that $\bar{\Phi}(\boldsymbol{\pi}) = \sum_{j\in M}\mathbb{E}[\bar{v}_j(C_j)]$ is affine in each coordinate $\pi_i(j)$ separately. Two consequences follow immediately. First, all same-agent second partials vanish, including the diagonal $\partial^2\bar{\Phi}/\partial\pi_i(j)^2 = 0$. Second, the summand $\mathbb{E}[\bar{v}_\ell(C_\ell)]$ depends on agent i only through $\pi_i(\ell)$, so $\partial^2\bar{\Phi}/\partial\pi_i(j)\partial\pi_k(\ell) = 0$ whenever $j \ne \ell$: the Hessian is block diagonal across tasks. The surviving entries are, for $i \ne k$ and $i, k \in N_j$,

$$H_{(i,j),(k,j)} = \mathbb{E}_{\mathbf{s}_{-\{i,k\}}}\big[v_j(B\cup\{i,k\}) - v_j(B\cup\{i\}) - v_j(B\cup\{k\}) + v_j(B)\big], \qquad B = C_j(\mathbf{s}_{-\{i,k\}}).$$

(b) Concavity of $\bar{\Phi}$ on $\boldsymbol{\Delta}$ is equivalent to $Q(\mathbf{w}) := \mathbf{w}^\top H\mathbf{w} \le 0$ for every *feasible* direction $\mathbf{w} \in \mathcal{T}$, i.e. every $\mathbf{w}$ with $\langle\mathbf{1}, w_i\rangle = 0$ for each i; this is the second-order characterisation on the affine hull of $\boldsymbol{\Delta}$. By (a) there is no within-agent block, so Q vanishes on each single-agent subspace and

$$Q(\mathbf{w}) \;=\; \sum_{i\ne k} w_i^\top H^{ik} w_k, \qquad H^{ik} := \big(H_{(i,j),(k,\ell)}\big)_{j\in S_i,\ \ell\in S_k},$$

which is *bilinear* in each pair (w_i, w_k). Replacing w_k by $-w_k$ (still feasible) negates the corresponding term, so $Q \le 0$ on $\mathcal{T}$ forces $w_i^\top H^{ik} w_k = 0$ for every $i \ne k$ and every feasible pair, i.e. $\mathsf{P}_i H^{ik}\mathsf{P}_k = 0$ for all $i \ne k$, i.e. $\mathsf{P}H\mathsf{P} = 0$, i.e. $Q \equiv 0$ on $\mathcal{T}$ and $\bar{\Phi}$ is affine on $\boldsymbol{\Delta}$. Conversely an affine function is concave. Note what this does *not* say: it does not say $H \equiv 0$. An agent with $|S_i| = 1$ contributes $w_i = 0$ to every feasible direction, so her rows of H are unconstrained; Remark 1 gives an instance where $\bar{\Phi}$ is affine on $\boldsymbol{\Delta}$ while the ambient H has a nonzero cross-partial, so the ambient form of the statement is false.

(c) Let $i \neq k$ and $j, j' \in S_i \cap S_k$ with $j \neq j'$, and put $w_i = w_k = e_j - e_{j'}$ and $w_r = 0$ otherwise; this is feasible. By (a),

$$Q(\mathbf{w}) \;=\; 2\sum_a H_{(i,a),(k,a)}\, w_i(a) w_k(a) \;=\; 2\big(H_{(i,j),(k,j)} + H_{(i,j'),(k,j')}\big).$$

If that sum is nonzero then either $Q(\mathbf{w}) > 0$, or Q evaluated at the feasible direction obtained by flipping the sign of w_k is positive; either way $\bar{\Phi}$ is not concave on $\mathbf{\Delta}$. Under submodularity each summand is ≤ 0 by (d), so the sum is nonzero exactly when at least one of the two cross-partials is.

(d) If each v_j is submodular then every displayed bracket is ≤ 0, so $H_{(i,j),(k,j)} \leq 0$. For $\mathbf{w} \geq 0$, $\mathbf{w}^\top H \mathbf{w} = \sum_j \sum_{i \neq k} H_{(i,j),(k,j)} w_{i,j} w_{k,j} \leq 0$. This is precisely continuous DR-submodularity ("up-concavity") in the sense of Bian et al. (2017); Calinescu et al. (2011). □

A worked counterexample. Take $n = m = 2$, $S_1 = S_2 = \{1, 2\}$, and $v_j(C) = f(|C|)$ for both tasks with $f(0) = 0$, $f(1) = 1$, $f(2) = 1.5$. This f is strictly concave, strictly increasing, and v_j is monotone submodular. Writing $\pi_1 = (p, 1-p)$ and $\pi_2 = (q, 1-q)$, the profile is split with probability $p(1-q) + (1-p)q$ and concentrated otherwise, so

$$\bar{\Phi}(p, q) = 1.5\big[pq + (1-p)(1-q)\big] + 2\big[p(1-q) + (1-p)q\big] = 1.5 + 0.5\,(p + q - 2pq),$$

whose Hessian $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$ has eigenvalues ± 1. Note what this rules out: "strictly concave anonymous values" is the most natural sufficient condition one would propose for concavity of $\bar{\Phi}$, and it fails on the smallest nontrivial instance. So does "non-positive off-diagonals", which combined with a zero diagonal is emphatically *not* negative semidefiniteness. Disjoint feasible sets $S_i \cap S_k = \emptyset$ do give concavity, but only by emptying the cross-block on feasible coordinates, a game in which no agent's choice affects any other, i.e. n independent single-agent problems. So does giving an agent no choice at all (Remark 1). Both are the same phenomenon: the absence of strategic interaction along directions the agents can move in.

Analysis without concavity. Concavity is conventionally used in stochastic Frank–Wolfe to convert the Frank–Wolfe gap into an optimality gap and thereby obtain a contraction $h_{t+1} \leq (1 - \eta_t) h_t + \cdots$, which is what allows a *decaying* step size $\eta_t \asymp 1/t$ Mokhtari et al. (2018); Zhang et al. (2020). We do not need that conversion. What we need is control of $\frac{1}{T}\sum_t \mathsf{FW}(\boldsymbol{\pi}^t)$, and by Lemma 3 that quantity *is* average exploitability, with no optimality-gap detour. It is obtained from L-smoothness plus $0 \leq \bar{\Phi} \leq m v_{\max}$ by telescoping (Step 4 of the proof of Theorem 1), which holds for any smooth bounded function. What losing the contraction costs is the *optimality-gap* route to a rate; we use a constant step size in its place, which is what caps the exponents we prove (Remark 7). We stress that a constant step size is our choice of analysis, not something the criterion forces: summation by parts accommodates nonincreasing η_t for an unweighted sum, and Dong et al. (2024) use decreasing steps for exactly this criterion.

B.5 Per-round computational and representation costs

There is no exponential general-versus-anonymous separation in `OSD-FW`'s per-round complexity. This section states what is true instead.

Polynomial per-round cost. For a general identity-dependent FVG, computing the unilateral-oracle estimator (5) for agent i requires evaluating $v_j(C_j(j, \mathbf{s}_{-i}))$ and $v_j(C_j(j, \mathbf{s}_{-i}) \setminus \{i\})$ for each $j \in S_i$: $\leq 2m$ oracle calls and $O(m)$ arithmetic. The drift correction (11) costs $O(mn)$ oracle calls and $O(mn)$ arithmetic (Proposition 2). The state carried between rounds is the momentum vector $\mathbf{d}_i^t \in \mathbb{R}^{|S_i|}$, the previous vertex, and the current policy: $O(m)$ numbers, plus the $O(n)$ announcements of the current round. *Nothing in Algorithm 1 scales exponentially in* n, on any instance, and no step of the regret analysis assumes otherwise.

The reason is that the algorithm never *represents* v_j. It issues queries at points determined by the realized and announced profiles, consumes each response once, and discards it. Caching responses across rounds would create a large table; such caching is an optional implementation choice for sample reuse. Our implementation and bounds do not use it. Conflating "the space of coalitions the algorithm could in principle visit over T rounds" with "memory the algorithm must allocate" is precisely the error to avoid here.

Exponential representation cost. Two genuinely exponential facts about the model class survive, and neither is a per-round cost of OSD−FW.

(i) *Representation.* $v_j : 2^{N_j} \to [0, v_{\max}]$ has $2^{|N_j|}$ degrees of freedom, and the monotone submodular functions on 2^{N_j} number $2^{2^{\Theta(n)}}$. Any *model-based* method, one that tabulates $\bar{v}_j$, regresses on a tabular basis, or maintains posterior beliefs over v_j, pays $\Omega(2^n)$ memory per task in the worst case. This is exactly the cost anonymity removes, collapsing the model to $m(n+1)$ parameters and making the ridge estimator (21) of Section A.8 possible. It is a statement about a family of model-based algorithms. This representation cost motivates the anonymous specialization.

(ii) *Identification.* Determining v_j requires $\Omega(2^n)$ oracle calls, since values on distinct subsets are unconstrained beyond monotonicity and submodularity. Our regret bounds are polynomial precisely because the algorithm never identifies v_j.

B.6 Extending Variant B to pure bandit feedback

The drift correction (11) evaluates the potential at profiles built from *other* agents' announced actions a_k^t, b_k^t. Under pure bandit feedback agent i observes only her own realized payoff $u_i^t(\mathbf{s})$; she sees neither the announcements nor s_k, and cannot evaluate a counterfactual coalition that was not formed. The drift correction is therefore inaccessible in that model, and Variant B does not run there.

The natural workaround, importance-weighting a pure-bandit observation to simulate the counterfactual, requires reweighting by the probability that the joint profile $\mathbf{s}_{-\{i,k\}}$ occurs, which under $\boldsymbol{\pi}$ can be as small as $m^{-(n-2)}$. This gives a second-moment blow-up exponential in n *for that particular estimator*.

Why there is no exponential lower bound to conjecture here. One might conjecture that *any* decentralized algorithm observing only its own realized payoff requires $T = \exp(\Omega(n)) \cdot \mathrm{poly}(1/\varepsilon)$ rounds on some family of identity-dependent FVGs. Such a conjecture is false: Corollary 2 already exhibits such an algorithm, Variant A with forced exploration, attaining $\mathbb{E}[\mathsf{NG}] \le \varepsilon$ in $T = \mathcal{O}(n^6 m^4 v_{\max}^6 \varepsilon^{-6} + n^8 m^4 v_{\max}^4 \varepsilon^{-4})$ rounds on the *general* class, under the same normalization and assumptions, and Dong et al. (2024) do better still at $\mathcal{O}(T^{4/5})$. The $m^{-(n-2)}$ blow-up bounds the cost of one route to the *drift correction*, not the cost of learning an equilibrium, and Variant A never forms that estimator.

What is actually open. The correct question is about the *rate*, not about polynomiality:

Open Problem 1. *Is $\mathcal{O}(T^{2/3})$ Nash regret, equivalently $\mathcal{O}(\varepsilon^{-3})$ sample complexity, attainable on general identity-dependent FVGs under* pure bandit *feedback with constants polynomial in n and m? Corollary 2 gives $\mathcal{O}(T^{5/6})$ with polynomial constants; Theorem 2 gives $\mathcal{O}(T^{2/3})$ but needs the auxiliary-draw protocol. Closing or separating that gap, either by a drift estimator that survives own-payoff feedback, or by a lower bound against the pure-bandit model, is the cleanest problem this work leaves open.*

Remark 8. *Anonymous FVGs do not support identity-dependent coalition values by definition. The Anon. FVG rows in Table 1 give bounds for the specialised algorithm on this subclass, not for the general identity-dependent case, and are conditional on Assumption 2.*

Scope and limitations. We summarise what the results do and do not claim.

- *Attribution.* The non-convex one-sample Frank–Wolfe template that Theorem 2 runs on, unbiased momentum with an interpolated-path drift correction, constant step size, average-Frank–Wolfe-gap progress measure, uniformly random output iterate, $\mathcal{O}(\varepsilon^{-3})$, is Theorem 2 of Zhang et al. (2020) and is not claimed as a contribution. Ours are the estimator (11), which makes that template implementable by agents who do not know each other's policies, and Lemma 3, which converts the resulting stationarity bound into a Nash-regret bound.
- No concavity of $\bar{\Phi}$ is assumed, and by Proposition 1 none is available in any FVG with cross-agent interaction.

- Variant B requires joint-action observability *and* the auxiliary-draw protocol of Section 3.3.2, and its value-oracle queries are *not* confined to a bounded neighbourhood of the played allocation. Variant A under unilateral-oracle feedback requires joint-action observability. Only Variant A under pure bandit feedback is fully local.
- The rate exponents ($T^{3/4}$, $T^{5/6}$, $T^{2/3}$) are what the uniform averaging demanded by Nash regret permits; they are weaker than the $\sqrt{T}$ rates available for last-iterate or optimality-gap criteria under decaying step sizes, and we do not claim otherwise (Remark 7).
- Per-round computation, memory and estimator variance are all polynomial in n and m for general identity-dependent FVGs (Table 7). We claim *no* computational separation between the general and anonymous cases for `OSD-FW`; what anonymity buys is a representable model and better constants (Section B.5).
- The anonymous specialization (Corollary 5) is conditional on the tracking hypothesis Assumption 2. We specify the estimator (21) and quantify the target's per-round drift, but do not prove the tracking bound for it under Frank–Wolfe dynamics.
- $\mathbf{d}^t$ is *not* a conditionally unbiased estimator of $\nabla\bar{\Phi}(\boldsymbol{\pi}^t)$ in Variant B, and the analysis does not assume it is (Step 0 of the proof of Theorem 2); what the drift correction delivers is a conditionally mean-zero *increment*.
- All environments are synthetic and fully specified; none is dataset-derived. The crowdsourcing and MRTA environments borrow benchmark structure but are not replications. Every configuration is run over resampled instances crossed with algorithm seeds, so error bars include instance variability, but instance sizes are modest ($n \leq 12$) because exact Nash-gap evaluation costs $\Theta(nm2^n)$.
- Variant B has no Nash-regret advantage in the matched-budget experiment. It is fourth of nine on $\mathcal{R}^T/T$, behind cheaper methods that convert the same budget into $20\times$ more rounds (Table 6); its advantage is on the final iterate, which we do not analyse. It spends $O(mn)$ oracle calls per round against $O(m)$ for Variant A and $O(1)$ for the bandit baselines.
- Under *pure bandit* feedback our $\mathcal{O}(T^{5/6})$ is superseded by the $\mathcal{O}(T^{4/5})$ of Dong et al. (2024), whose framework covers FVGs after payoff normalisation, and which also beats our bandit variant empirically at long horizons (Table 6). The pure-bandit rate is included only as an ablation.
- The certified exponents *are* matched empirically once the quantity the theorems bound is the one fitted: the fitted slopes of $\mathcal{R}^T/T$ are -0.239, -0.335 and -0.151 against $-1/4$, $-1/3$ and $-1/6$ (Figure 1a). What converges much faster than certified is the *final iterate*, for which we prove nothing. This is evidence of consistency over the horizons tested, not of tightness or optimality; no matching lower bound is offered.

B.7 Corrections relative to the previous version

The body states the revised claims; each correction below points to where it is made.

1. **Prior work.** Dong et al. (2024) supersede our pure-bandit $\mathcal{O}(T^{5/6})$ with $\mathcal{O}(T^{4/5})$, and the Frank–Wolfe-gap-to-Nash-gap relation is theirs; both are cited, tabulated and used as a baseline, and neither is presented as a contribution.
2. **The protocol is not private.** The transcript determines every policy exactly; the claim is communication and query efficiency only (Remark 5).
3. **Step sizes.** Summation by parts handles nonincreasing η_t, so our exponents describe the constant-step family we analyse, not a barrier (Remark 7).
4. **Proposition 1.** Concavity is characterised on feasible directions ($\mathsf{P}H\mathsf{P} = 0$), not the ambient Hessian (Remark 1).
5. **The drift estimator.** The Jacobian–vector form against $\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}$ is implementable by no agent; the sampled construction of Section 3.3.2 replaces it at $O(mn)$ queries.
6. **Analysis fixes.** Pure bandit runs on the shrunken simplex at the played policy (Corollary 2); the drift term carries its aggregation factor n; oracle noise is drawn per call (Appendix B.3).
7. **Complexity.** No exponential general-versus-anonymous separation is claimed; the exponential is representational (Appendix B.5), and the anonymous specialization is conditional on Assumption 2.

8. **Experiments.** Metrics are exact and the primary column is $\mathcal{R}^T/T$; the matched-budget comparison (Table 6) shows `OSD-FW-B` does *not* lead on Nash regret, and we no longer claim it does.

C Directions for future work

This section states the open directions in enough detail to be acted on. Each begins from something the paper leaves unresolved rather than from a generic wish list, and we indicate where we believe the difficulty actually lies.

C.1 An analysis that can see the gradient path

The bias term in Lemma 7 is $10L^2D^2\eta^2T^{3/2}$, obtained by bounding each per-round gradient increment as $\|\nabla\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla\bar{\Phi}(\boldsymbol{\pi}^{t-1})\| \le L\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}\| \le LD\eta$ and summing without cancellation. Two things follow. First, the bound depends on the instance only through the worst-case smoothness constant L, so it assigns the same lag penalty to instances whose measured gradient-path coherence differs by an order of magnitude (Section 4). Second, the drift correction's contribution to the bound is a constant factor, whereas its measured contribution ranges from nothing to 16%. The theory, as it stands, cannot express the phenomenon the experiments isolate.

The repair we would attempt is to carry the realized increments rather than their worst-case surrogate. Write $D^t = \mathsf{P}(\nabla\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla\bar{\Phi}(\boldsymbol{\pi}^{t-1}))$ and $w_{t,s} = \prod_{r=s+1}^{t}(1-\rho_r)$, so that the lag of the biased recursion (6) is $\big\|\sum_{s\le t} w_{t,s}D^s\big\|$ up to the martingale part. The present proof passes to $\sum_s w_{t,s}\|D^s\| \le LD\eta\sum_s w_{t,s}$, which discards both the realized magnitudes and any cancellation. Keeping the realized magnitudes alone would already make the bound instance-dependent, and it is the version consistent with our measurements: the momentum estimate is an average over a window of roughly $t^{2/3}$ rounds, so what hurts it is local path length inside that window, not net displacement. On an oscillating instance the average sits near the middle of the swing while the Frank–Wolfe oracle needs the current gradient, and the drift term is exactly what restores it; on a coherent instance the window average is already close to the current value and the correction has little to add. Gradient-path coherence, being the ratio of net displacement to path length, is then a proxy for the operative quantity rather than the quantity itself.

We flag one trap. The more obvious refinement — bounding the lag by the net displacement $\|\sum_s w_{t,s}D^s\|$ and concluding that the correction matters when that is large — predicts the *opposite* of the measured ordering, since the identity-dependent families are the coherent ones. Any proposed bound should therefore be checked against Figure 1b before it is believed.

Open Problem 2. *Exhibit a functional Ψ of the realized gradient path for which the Variant A lag is $\Theta(\Psi)$ and the Variant B lag is $o(\Psi)$, and which is large on the anonymous instances of Section 4 and small on the identity-dependent ones. Local path length within the momentum window is our candidate; net displacement is not one.*

C.2 Last-iterate guarantees

Every rate in this paper controls $\frac{1}{T}\sum_t \mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)]$, and the experiments repeatedly show the final iterate to be one to three orders of magnitude better than that average — the one metric on which Variant B leads is one the theory does not cover. Closing this is not cosmetic: a practitioner deploys the last policy, not a uniformly sampled one.

The structure to exploit is that $\bar{\Phi}$ is an exact potential, so the natural argument is ascent rather than descent. Smoothness and the Frank–Wolfe step give

$$\mathbb{E}[\bar{\Phi}(\boldsymbol{\pi}^{t+1})] \;\ge\; \mathbb{E}[\bar{\Phi}(\boldsymbol{\pi}^t)] + \eta\,\mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)] - \eta D\,\mathbb{E}\|e^t\| - \tfrac{LD^2\eta^2}{2},$$

so $\bar{\Phi}(\boldsymbol{\pi}^t)$ is a submartingale up to terms that are summable at the paper's schedules. Since $\bar{\Phi}$ is bounded by $mv_{\max}$, this recovers the average bound immediately; what it does not give is that $\mathsf{FW}(\boldsymbol{\pi}^t)$ stays small once it becomes small. The missing ingredient is a stability statement — that the gap cannot rise substantially after it has fallen — which for a constant step size is false in general and will need either a decreasing η_t or a local argument near the equilibrium set. A promising

intermediate target is a bound on $\mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^\tau)]$ for τ uniform on the *last* $T/2$ rounds, which is already much closer to practice and which the submartingale property may reach without a new idea.

C.3 Decreasing schedules under oracle feedback

Our exponents are established for a constant η, and Dong et al. (2024) obtain their better pure-bandit exponent with a decreasing step size. Whether the oracle-feedback exponents $3/4$ and $2/3$ are similarly improvable is open, and the question is concrete: in the error recursion behind Lemma 7, put $\eta_t = (t+1)^{-a}$ and $\rho_t = (t+1)^{-p}$ and optimize the resulting frontier in (a, p) rather than fixing p to $1/2$ or $2/3$ as we do. Summation by parts already permits nonincreasing η_t for the unweighted sum, so no obstruction is known; what is missing is the optimization itself and a matching instance. Our own ablation is a caution rather than an encouragement here: the entire measured gap between Variants A and B was a step-size effect, so a schedule study may well move the empirical numbers more than any estimator change.

C.4 Lower bounds, of which this paper has none

We show consistency between fitted and certified exponents, not optimality, and nothing here rules out a better rate for any of the three feedback models. Three targets, in increasing order of difficulty. (i) A query-complexity lower bound: how many value-oracle calls are needed to certify an ε-approximate equilibrium of an identity-dependent FVG, independent of the algorithm? The $2^{\Theta(n)}$ representation cost of the class (Appendix B.5) does not by itself force a large query count, and separating the two is the cleanest question. (ii) A separation between the auxiliary-draw and pure-bandit models, which would justify the protocol rather than merely describing it; Open Problem 1 is the rate version of this. (iii) Optimality of $T^{2/3}$ for one-sample projection-free methods on nonconvex multilinear extensions, which would settle whether the template inherited from Zhang et al. (2020) is the right one.

C.5 Removing the tracking assumption

The anonymous rows of Table 1 are conditional on Assumption 2 for the windowed ridge estimator (21), and a conditional row is worth much less than an unconditional one. We believe the assumption is provable rather than fundamental, because Frank–Wolfe controls its own drift: the iterates satisfy $\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}\| \le D\eta$ by construction, so the regression target moves at a rate the algorithm itself sets. A window of W rounds then carries bias $O(D\eta W)$ from target movement and variance $O(dv_{\max}^2/W)$ with $d = m(n+1)$, and balancing the two gives a tracking error that vanishes with η. Turning this into a proof requires handling the dependence between the window's design points and the policy path, which is where a self-normalized martingale bound would enter.

C.6 A protocol that buys privacy as well as bandwidth

The auxiliary-draw protocol is deliberately not private: the announced vertices, the step size and the prescribed initialization reconstruct every policy (Remark 5), and we have been careful not to claim otherwise. It is natural to ask what it would cost to make it private. Replacing the announcement a_k^t by a randomized response over S_k at parameter $\varepsilon_{\mathrm{dp}}$ keeps the draw unbiased after the standard debiasing, at the price of inflating the drift variance σ_H^2 by a factor $\big((e^{\varepsilon_{\mathrm{dp}}} + m - 1)/(e^{\varepsilon_{\mathrm{dp}}} - 1)\big)^2$. Because σ_H enters Theorem 2 only through the constant, the $T^{2/3}$ exponent should survive, with a constant growing like $m/\varepsilon_{\mathrm{dp}}$ in the high-privacy regime. We state this as a conjecture rather than a result: the debiasing interacts with the coin $c_k^t \sim \mathrm{Bern}(\alpha^t\eta)$, and it is not obvious that the composed estimator stays conditionally unbiased for the gradient increment.

Open Problem 3. *Does Variant B retain $\mathcal{O}(T^{2/3})$ Nash regret when the announcements a_k^t are released under local differential privacy at parameter $\varepsilon_{\mathrm{dp}}$, with constants polynomial in n, m and $1/\varepsilon_{\mathrm{dp}}$? The vertex announcements b_k^t are the harder half, being deterministic functions of an agent's own linear oracle.*

C.7 Query cost that falls as play concentrates

The sampled estimator (11) skips every summand with $a_k^t = b_k^t$, an event of probability $\pi_k^t(b_k^t)$ that tends to one as the policies concentrate. Its realized cost is therefore $O\big(m\sum_k(1-\pi_k^t(b_k^t))\big)$ rather than $O(mn)$, which is why it beats the exact Jacobian–vector implementation under a matched budget despite the latter's lower variance (Appendix A.7). We conjecture this can be made a theorem. If the linear oracle settles on a fixed vertex after some round, the Frank–Wolfe contraction gives $1-\pi_k^t(b_k) = O((1-\eta)^t)$ and hence $O(mn/\eta) = O(mnT^{2/3})$ total oracle calls over the horizon, against $O(mnT)$ for the worst case: a $T^{1/3}$ saving, and an anytime guarantee rather than a worst-case one. The obstruction is that μ_k^t need not settle — on the oscillating instances of Section 4 it demonstrably does not, and those are exactly the instances where the drift correction is worth having. A cost bound that interpolates between the two regimes is the useful statement, and it would likely use the same path functional as Open Problem 2.

C.8 Beyond the model

Three extensions look tractable, in the order given. *Multi-task assignment*: allowing each agent a set of tasks subject to a capacity constraint replaces the product of simplices by a product of matroid polytopes, on which the Frank–Wolfe linear oracle is still a greedy computation; Lemma 3 is the step that needs re-proving, since its decoupling uses the simplex structure. *Non-monotone values*: our smoothness constant and the potential range both use monotonicity, though the experiments already run on a non-monotone environment outside Definition 1 without qualitative change (Appendix A.7); making that rigorous mostly means re-deriving Lemma 4. *Non-stationary values*: if $\bar{v}_j$ drifts over time, Nash regret must be measured against a moving equilibrium, and the natural target is a dynamic-regret bound in terms of the path length of the equilibrium sequence — which connects back to the same path-functional question that Open Problem 2 poses for the gradient.

D Proofs

D.1 Proof of Lemma 1

Proof. Fix agent i, profile $\mathbf{s}$, and alternative $s_i' \neq s_i$. Let $j = s_i$, $j' = s_i'$. When i switches from j to j': $C_j(s_i', \mathbf{s}_{-i}) = C_j(\mathbf{s}) \setminus \{i\}$, $C_{j'}(s_i', \mathbf{s}_{-i}) = C_{j'}(\mathbf{s}) \cup \{i\}$, and all other coalitions are unchanged. Therefore,

$$\begin{aligned}\Phi(s_i', \mathbf{s}_{-i}) - \Phi(\mathbf{s}) &= \big[v_{j'}(C_{j'}(\mathbf{s}) \cup \{i\}) - v_{j'}(C_{j'}(\mathbf{s}))\big] + \big[v_j(C_j(\mathbf{s}) \setminus \{i\}) - v_j(C_j(\mathbf{s}))\big] \\ &= \delta_{i,j'}(C_{j'}(\mathbf{s}) \cup \{i\}) - \delta_{i,j}(C_j(\mathbf{s})) = u_i(s_i', \mathbf{s}_{-i}) - u_i(\mathbf{s}).\end{aligned}$$

□

Corollary 6. *Every FVG has a pure Nash equilibrium (any maximizer of Φ over the finite set $\mathcal{S}$).*

D.2 Proof of Lemma 2

Proof. By independence of agents and multilinearity of $\bar{\Phi}$: $\partial\bar{\Phi}/\partial\pi_i(j) = \mathbb{E}_{\mathbf{s}_{-i}}[\Phi(j, \mathbf{s}_{-i})]$. Fix reference $j_0 \in S_i$. By Lemma 1, $\Phi(j, \mathbf{s}_{-i}) = \Phi(j_0, \mathbf{s}_{-i}) + u_i(j, \mathbf{s}_{-i}) - u_i(j_0, \mathbf{s}_{-i})$. Taking expectation and setting $c_i(\boldsymbol{\pi}_{-i}) := \mathbb{E}_{\mathbf{s}_{-i}}[\Phi(j_0, \mathbf{s}_{-i}) - u_i(j_0, \mathbf{s}_{-i})]$, which is independent of j and satisfies $|c_i| \leq \mathbb{E}[|\Phi|] + \mathbb{E}[|u_i|] \leq (m+1)v_{\max}$, gives the identity. □

D.3 Proof of Lemma 3

Proof. Since $\boldsymbol{\Delta}$ is a product of simplices, $\mathsf{FW}(\boldsymbol{\pi}) = \sum_i \max_{\mu_i}\langle\nabla_{\pi_i}\bar{\Phi}, \mu_i - \pi_i\rangle$. By Lemma 2, $\nabla_{\pi_i}\bar{\Phi} = \nabla_{\pi_i}U_i + c_i\mathbf{1}$; since $\mu_i, \pi_i \in \Delta(S_i)$, the constant cancels. The maximum is achieved at the point mass on $\arg\max_j \mathbb{E}_{\mathbf{s}_{-i}}[u_i(j, \mathbf{s}_{-i})]$, giving $U_i^\star(\boldsymbol{\pi}_{-i}) - U_i(\boldsymbol{\pi})$ for each i. Summing gives $\mathsf{FW} = \sum_i(U_i^\star - U_i)$; non-negativity of each term yields the sandwich bound. □

Proof of Corollary 1. $\mathsf{FW}(\boldsymbol{\pi}^\star) = 0$ iff every summand $U_i^\star(\boldsymbol{\pi}_{-i}^\star) - U_i(\boldsymbol{\pi}^\star) = 0$, i.e., no agent has a profitable deviation. This is the Nash condition and holds regardless of concavity. □

D.4 Proof of Lemma 4

Proof. By multilinearity, the only nonzero cross-partial entries are $\partial^2\bar{\Phi}/\partial\pi_i(j)\partial\pi_k(\ell)$ for $i \neq k$, $j = \ell, i, k \in N_j$, equal to $\mathbb{E}_{\mathbf{s}_{-\{i,k\}}}[v_j(B\cup\{i,k\}) - v_j(B\cup\{i\}) - v_j(B\cup\{k\}) + v_j(B)]$, bounded in absolute value by $2v_{\max}$ by submodularity. Same-agent partials vanish by multilinearity. For any unit vector $\mathbf{v}$:

$$|\mathbf{v}^\top H\mathbf{v}| \le 2v_{\max}\sum_{j\in M}\sum_{\substack{i\neq k\\ i,k\in N_j}}|v_{i,j}|\,|v_{k,j}| \le 2v_{\max}\sum_{j\in M}(n-1)\sum_{i\in N_j}v_{i,j}^2 \le 2(n-1)mv_{\max}\|\mathbf{v}\|^2,$$

where the second inequality uses AM-GM and the third uses $\sum_{i,j}v_{i,j}^2 \le \|\mathbf{v}\|^2$. Thus $\|\nabla^2\bar{\Phi}\|_{\mathrm{op}} \le L = 2(n-1)mv_{\max}$. □

D.5 Proof of Lemma 5

Proof. Component j: $\mathbb{E}_{\mathbf{s}_{-i}}[\delta_{i,j}(C_j(j,\mathbf{s}_{-i}))] = \mathbb{E}_{\mathbf{s}_{-i}}[u_i(j,\mathbf{s}_{-i})] = \partial U_i(\boldsymbol{\pi}^t)/\partial\pi_i(j) = \partial\bar{\Phi}(\boldsymbol{\pi}^t)/\partial\pi_i(j) - c_i$; the constant shift is absorbed by the FW oracle. □

D.6 Proof of Lemma 7

Proof. Work with the stacked projected error $e^t = \mathsf{P}(\mathbf{d}^t - \nabla\bar{\Phi}(\boldsymbol{\pi}^t))$, where $\mathsf{P} = \mathrm{diag}(\mathsf{P}_1,\dots,\mathsf{P}_n)$ (Appendix B.2). Applying P to (6),

$$e^t = (1-\rho_t)\,e^{t-1} + \underbrace{(1-\rho_t)\mathsf{P}\big(\nabla\bar{\Phi}(\boldsymbol{\pi}^{t-1}) - \nabla\bar{\Phi}(\boldsymbol{\pi}^t)\big)}_{=:b^t} + \rho_t\,\mathsf{P}\xi^t, \qquad \xi^t := \widehat{\nabla}U(\boldsymbol{\pi}^t) - \nabla U(\boldsymbol{\pi}^t), \tag{24}$$

where we used $\mathsf{P}(c\mathbf{1}) = 0$ to replace $\bar{\Phi}$ by U in the noise term. By construction $\mathbb{E}[\mathsf{P}\xi^t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}\|\mathsf{P}\xi^t\|^2 \le n\tilde{\sigma}^2$. Since $\|\mathsf{P}\|_{\mathrm{op}} \le 1$ and $\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}\| = \eta\|\boldsymbol{\mu}^{t-1} - \boldsymbol{\pi}^{t-1}\| \le \eta D$, L-smoothness gives $\|b^t\| \le L\eta D$ deterministically.

Step 1 (closing the recursion). Conditioning on $\mathcal{F}_{t-1}$ makes $(1-\rho_t)e^{t-1} + b^t$ deterministic, so the cross term with $\rho_t\mathsf{P}\xi^t$ vanishes:

$$\mathbb{E}\|e^t\|^2 = \mathbb{E}\big\|(1-\rho_t)e^{t-1} + b^t\big\|^2 + \rho_t^2\,\mathbb{E}\|\mathsf{P}\xi^t\|^2.$$

For any $\beta > 0$, $\|x+y\|^2 \le (1+\beta)\|x\|^2 + (1+\beta^{-1})\|y\|^2$. Take $\beta = \rho_t/2$. For $\rho_t \in [0,1]$,

$$(1+\tfrac{\rho_t}{2})(1-\rho_t)^2 = 1 - \tfrac{3\rho_t}{2} + \tfrac{\rho_t^3}{2} \le 1-\rho_t, \qquad (1+\tfrac{2}{\rho_t})(1-\rho_t)^2 \le 1 + \tfrac{2}{\rho_t} \le \tfrac{3}{\rho_t},$$

whence

$$\mathbb{E}\|e^t\|^2 \;\le\; (1-\rho_t)\,\mathbb{E}\|e^{t-1}\|^2 \;+\; \frac{3}{\rho_t}L^2D^2\eta^2 \;+\; \rho_t^2\,n\tilde{\sigma}^2. \tag{25}$$

This is where the argument must not use $(a+b+c)^2 \le 3(a^2+b^2+c^2)$: that bound leaves the coefficient $3(1-\rho_t)^2$, which exceeds 1 whenever $\rho_t < 2/3$, making (25) expansive and the induction below false.

Step 2 (induction). With $\rho_t = 2/(t+4)^{1/2}$ we have $3/\rho_t = \frac{3}{2}(t+4)^{1/2}$ and $\rho_t^2 = 4/(t+4)$. We claim $\mathbb{E}\|e^t\|^2 \le Q/(t+5)^{1/2}$ for $Q = 3n\tilde{\sigma}^2 + 10L^2D^2\eta^2T^{3/2} + 3nmv_{\max}^2$. For the base case, $\mathbf{d}^0 = 0$ gives $\|e^0\| = \|\mathsf{P}\nabla U(\boldsymbol{\pi}^1)\|$, whose per-agent squared norm is at most $mv_{\max}^2$ (each of the $\le m$ entries of $\nabla_{\pi_i}U_i$ lies in $[0, v_{\max}]$ and projection is a contraction), so $\mathbb{E}\|e^0\|^2 \le nmv_{\max}^2 \le Q/\sqrt{5}$. Assume $\mathbb{E}\|e^{t-1}\|^2 \le Q/(t+4)^{1/2}$. Then

$$\mathbb{E}\|e^t\|^2 \le \frac{Q}{(t+4)^{1/2}} - \frac{2Q}{t+4} + \tfrac{3}{2}(t+4)^{1/2}L^2D^2\eta^2 + \frac{4n\tilde{\sigma}^2}{t+4}.$$

Since $(t+5)^{1/2} - (t+4)^{1/2} \le \frac{1}{2}(t+4)^{-1/2}$,

$$\frac{Q}{(t+4)^{1/2}} - \frac{Q}{(t+5)^{1/2}} \le \frac{Q}{2(t+4)^{3/2}} \le \frac{Q}{2\sqrt{5}\,(t+4)} \le \frac{0.224\,Q}{t+4},$$

so it suffices that $\frac{3}{2}(t+4)^{1/2}L^2D^2\eta^2 + 4n\tilde{\sigma}^2/(t+4) \le 1.776\,Q/(t+4)$, i.e.

$$\tfrac{3}{2}(t+4)^{3/2}L^2D^2\eta^2 + 4n\tilde{\sigma}^2 \;\le\; 1.776\,Q.$$

For $1 \le t \le T$ we have $(t+4)^{3/2} \le (5T)^{3/2} \le 11.2\,T^{3/2}$, so the left side is at most $16.8\,L^2D^2\eta^2T^{3/2} + 4n\tilde{\sigma}^2$, while $1.776\,Q \ge 17.8\,L^2D^2\eta^2T^{3/2} + 5.3\,n\tilde{\sigma}^2$. The induction closes.

Step 3 (T-freeness). Q is free of T exactly when $\eta^2T^{3/2} = O(1)$, i.e. $\eta = O(T^{-3/4})$. Taking $\eta = T^{-3/4}$ gives $Q = 3n\tilde{\sigma}^2 + 10L^2D^2 + 3nmv_{\max}^2$. Note the contrast with the schedule $(\eta, \rho_t) = (T^{-2/3}, 4(t+8)^{-2/3})$: there the analogous requirement is $\eta^2T^2 = O(1)$, which $\eta = T^{-2/3}$ violates by a factor $T^{2/3}$, so Q would carry a $T^{2/3}$ and Step 5 of Theorem 1 would degrade from $T^{-1/4}$ to $T^{-1/6}$. □

D.7 Proof of Lemma 6

Proof. Write $\mathbf{J}_i(\boldsymbol{\pi})_{j,(k,\ell)} := \partial^2\bar{\Phi}(\boldsymbol{\pi})/\partial\pi_i(j)\partial\pi_k(\ell)$. Because $\bar{\Phi}$ is multilinear,

$$\mathbf{J}_i(\boldsymbol{\pi})_{j,(k,\ell)} = \mathbb{E}_{\mathbf{s}_{-\{i,k\}}\sim\boldsymbol{\pi}_{-\{i,k\}}}\big[\Phi(j,\ell,\mathbf{s}_{-\{i,k\}})\big] \quad (k \ne i), \qquad \mathbf{J}_i(\boldsymbol{\pi})_{j,(i,\ell)} = 0. \tag{26}$$

Step 1 (the assembled profile has the right law). By the Frank–Wolfe update, $\pi_k^t = (1-\eta)\pi_k^{t-1} + \eta\mu_k^{t-1}$, hence

$$\pi_k^t(\alpha) := \pi_k^{t-1} + \alpha(\pi_k^t - \pi_k^{t-1}) = (1-\alpha\eta)\,\pi_k^{t-1} + \alpha\eta\,\mu_k^{t-1}.$$

By construction (10), $\mathbf{s}'^t_k$ equals b_k^t with probability $\alpha\eta$ and a_k^t otherwise, where $b_k^t \sim \mu_k^{t-1}$ (a point mass) and $a_k^t \sim \pi_k^{t-1}$. Therefore $\mathbf{s}'^t_k \sim \pi_k^t(\alpha)$ conditionally on $(\mathcal{F}_{t-1}, \alpha)$, and by independence of the announcements across agents and of the coins across k given α, the coordinates $\{\mathbf{s}'^t_k\}_{k\ne i}$ are mutually independent. Hence $\mathbf{s}'^t_{-i} \sim \boldsymbol{\pi}^t(\alpha)_{-i}$.

Step 2 (each summand is unbiased for its column block). Fix $k \ne i$ and $j \in S_i$. The coordinates $\mathbf{s}'^t_{-\{i,k\}}$ are functions of $\{a_m^t, b_m^t, c_m^t\}_{m\notin\{i,k\}}$ only, and are therefore conditionally independent of (a_k^t, b_k^t) given $(\mathcal{F}_{t-1}, \alpha)$. Conditioning on $(\mathcal{F}_{t-1}, \alpha)$ and on $\mathbf{s}'^t_{-\{i,k\}}$, and using $b_k^t \sim \mu_k^{t-1}$, $a_k^t \sim \pi_k^{t-1}$,

$$\mathbb{E}\big[\Phi(j, b_k^t, \mathbf{s}'^t_{-\{i,k\}}) - \Phi(j, a_k^t, \mathbf{s}'^t_{-\{i,k\}})\big] = \sum_{\ell\in S_k}\Phi(j,\ell,\mathbf{s}'^t_{-\{i,k\}})\big(\mu_k^{t-1}(\ell) - \pi_k^{t-1}(\ell)\big).$$

Taking the outer expectation over $\mathbf{s}'^t_{-\{i,k\}}$ and using (26) at $\boldsymbol{\pi} = \boldsymbol{\pi}^t(\alpha)$, then multiplying by η and invoking $\eta(\mu_k^{t-1} - \pi_k^{t-1}) = \pi_k^t - \pi_k^{t-1}$,

$$\eta\,\mathbb{E}\big[\Phi(j, b_k^t, \mathbf{s}'^t_{-\{i,k\}}) - \Phi(j, a_k^t, \mathbf{s}'^t_{-\{i,k\}}) \;\big|\; \mathcal{F}_{t-1}, \alpha\big] = \sum_{\ell\in S_k}\mathbf{J}_i(\boldsymbol{\pi}^t(\alpha))_{j,(k,\ell)}\big(\pi_k^t(\ell) - \pi_k^{t-1}(\ell)\big).$$

Step 3 (sum and integrate). Summing over $k \ne i$, legitimate by linearity of expectation, notwithstanding that the summands are dependent through the shared coordinates of $\mathbf{s}'^t$, and noting that the $k = i$ block of $\mathbf{J}_i$ vanishes by (26), we get $\mathbb{E}[\tilde{\boldsymbol{\Delta}}_i^t \mid \mathcal{F}_{t-1}, \alpha] = \mathbf{J}_i(\boldsymbol{\pi}^t(\alpha))(\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1})$. Integrating over $\alpha \sim \mathrm{Unif}[0,1]$ and applying the fundamental theorem of calculus to the smooth path $\alpha \mapsto \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t(\alpha))$,

$$\mathbb{E}[\tilde{\boldsymbol{\Delta}}_i^t \mid \mathcal{F}_{t-1}] = \Big(\int_0^1 \mathbf{J}_i(\boldsymbol{\pi}^t(\alpha))\,d\alpha\Big)(\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}) = \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi}^{t-1}).$$

Step 4 (almost-sure bound and implementability). Each summand is a difference of potentials at two profiles differing only in coordinate k, which alters exactly two coalitions, so it lies in $[-2v_{\max}, 2v_{\max}]$ by Assumption 1; summing $n-1$ of them and multiplying by η gives $|[\tilde{\boldsymbol{\Delta}}_i^t]_j| \le 2\eta(n-1)v_{\max}$, and $\|\tilde{\boldsymbol{\Delta}}_i^t\| \le 2\eta(n-1)\sqrt{m}\,v_{\max}$ since $|S_i| \le m$. Finally, inspect what enters the right-hand side of (11): the announced actions $\{a_k^t, b_k^t\}_{k\ne i}$, the public scalar η, agent i's own α^t and coins, and value-oracle responses. No policy, policy increment, or distribution belonging to another agent appears. □

Remark 9 (Why the centralized form could not be used)**.** *The estimator (11) is not a rewriting of* $\widehat{\mathbf{J}}_i(\boldsymbol{\pi}^t(\alpha);\mathbf{s}')(\boldsymbol{\pi}^t-\boldsymbol{\pi}^{t-1})$ *with better constants; it is a different estimator with the same conditional mean. The Jacobian–vector form needs two objects no agent has: the joint increment* $\boldsymbol{\pi}^t-\boldsymbol{\pi}^{t-1}$*, which it multiplies explicitly, and a sample from* $\boldsymbol{\pi}^t(\alpha)$*, which is an interpolation of policies agent* i *cannot see. Step 1 above obtains the second from other agents'* samples*, and Step 2 obtains the first from the identity* $\eta(\mu_k^{t-1}-\pi_k^{t-1})=\pi_k^t-\pi_k^{t-1}$ *together with a* sample *from each of* μ_k^{t-1} *and* π_k^{t-1}*, again drawn by their owner. That is the whole construction depends on the Frank–Wolfe identity expressing the increment as a difference of two distributions.*

D.8 Proof of Theorem 1

Proof. Throughout, $\eta_t\equiv\eta=T^{-3/4}$, $\boldsymbol{\mu}^t$ is the linear-oracle output, and $e^t=\mathsf{P}(\mathbf{d}^t-\nabla\bar{\Phi}(\boldsymbol{\pi}^t))$ is the stacked projected error of Lemma 7. *No step uses concavity of* $\bar{\Phi}$; we flag at each stage what is used instead.

Step 1 (Smoothness ascent). By L-smoothness of $\bar{\Phi}$ (Lemma 4),

$$\bar{\Phi}(\boldsymbol{\pi}^{t+1})\geq\bar{\Phi}(\boldsymbol{\pi}^t)+\langle\nabla\bar{\Phi}(\boldsymbol{\pi}^t),\boldsymbol{\pi}^{t+1}-\boldsymbol{\pi}^t\rangle-\tfrac{L}{2}\|\boldsymbol{\pi}^{t+1}-\boldsymbol{\pi}^t\|^2.$$

Smoothness alone; a smooth non-concave function satisfies this.

Step 2 (Bounding the linear term). Write $\boldsymbol{\pi}^{t+1}-\boldsymbol{\pi}^t=\eta(\boldsymbol{\mu}^t-\boldsymbol{\pi}^t)$ with $\|\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\|\leq D$. Because $\boldsymbol{\mu}^t-\boldsymbol{\pi}^t$ is a stacked difference of points of the simplices, every inner product below may be taken against projected vectors:

$$\langle\nabla\bar{\Phi},\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\rangle=\langle\mathsf{P}\nabla\bar{\Phi},\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\rangle=\langle\mathsf{P}\mathbf{d}^t,\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\rangle-\langle e^t,\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\rangle.$$

Let $\boldsymbol{\mu}^\star\in\arg\max_{\boldsymbol{\mu}\in\boldsymbol{\Delta}}\langle\nabla\bar{\Phi}(\boldsymbol{\pi}^t),\boldsymbol{\mu}\rangle$. Since $\boldsymbol{\mu}^t$ maximizes $\langle\mathbf{d}^t,\cdot\rangle$ (equivalently $\langle\mathsf{P}\mathbf{d}^t,\cdot\rangle$) over $\boldsymbol{\Delta}$, $\langle\mathsf{P}\mathbf{d}^t,\boldsymbol{\mu}^t-\boldsymbol{\pi}^t\rangle\geq\langle\mathsf{P}\mathbf{d}^t,\boldsymbol{\mu}^\star-\boldsymbol{\pi}^t\rangle$, and Cauchy–Schwarz on the two error terms gives

$$\langle\nabla\bar{\Phi}(\boldsymbol{\pi}^t),\boldsymbol{\pi}^{t+1}-\boldsymbol{\pi}^t\rangle\;\geq\;\eta\,\mathsf{FW}(\boldsymbol{\pi}^t)-2\eta D\,\|e^t\|. \tag{27}$$

Only the geometry of $\boldsymbol{\Delta}$ and the definition of the linear oracle are used.

Step 3 (Rearrange). Combining Steps 1–2, $\eta\,\mathsf{FW}(\boldsymbol{\pi}^t)\leq\bar{\Phi}(\boldsymbol{\pi}^{t+1})-\bar{\Phi}(\boldsymbol{\pi}^t)+2\eta D\|e^t\|+\frac{L\eta^2D^2}{2}$.

Step 4 (Sum and telescope). Summing over $t=1,\dots,T$, the first difference telescopes. Since $\Phi(\mathbf{s})\leq mv_{\max}$ for every $\mathbf{s}$ (the potential is a sum of m task values each bounded by $v_{\max}$; Appendix B.1), $0\leq\bar{\Phi}\leq mv_{\max}$ and $\bar{\Phi}(\boldsymbol{\pi}^{T+1})-\bar{\Phi}(\boldsymbol{\pi}^1)\leq mv_{\max}$. Dividing by ηT,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)]\leq\frac{mv_{\max}}{\eta T}+\frac{2D}{T}\sum_{t=1}^{T}\mathbb{E}\|e^t\|+\frac{L\eta D^2}{2}. \tag{28}$$

This is the step at which a concavity assumption is conventionally invoked, to argue that Frank–Wolfe stationary points are global maximizers. It is not needed: the telescope bounds the total *increase* of $\bar{\Phi}$ by its range, which holds for any bounded function whatever its curvature. All that is required is Assumption 1. We take η constant, which makes the left-hand side the uniform average directly. This is a convenience, not a necessity: for nonincreasing η_t the same telescope runs through summation by parts, giving $\sum_t(\bar{\Phi}^{t+1}-\bar{\Phi}^t)/\eta_t\leq mv_{\max}/\eta_T$ because $1/\eta_{t-1}-1/\eta_t\leq 0$ and $\bar{\Phi}\geq 0$. We do not pursue the decreasing-step analysis (Remark 7).

Step 5 (Error sum). By Jensen and Lemma 7, $\mathbb{E}\|e^t\|\leq\sqrt{Q}\,(t+5)^{-1/4}$. Since $\sum_{t=1}^{T}(t+5)^{-1/4}\leq\frac{4}{3}(T+5)^{3/4}\leq\frac{4}{3}\cdot 2^{3/4}T^{3/4}\leq 2.25\,T^{3/4}$ for $T\geq 5$,

$$\frac{2D}{T}\sum_{t=1}^{T}\mathbb{E}\|e^t\|\;\leq\;\frac{4.5\,D\sqrt{Q}}{T^{1/4}}.$$

Step 6 (Final bound). Substituting $\eta=T^{-3/4}$ into (28) gives $mv_{\max}T^{-1/4}+5D\sqrt{Q}\,T^{-1/4}+\frac{LD^2}{2}T^{-3/4}$, which is (13). With $\sqrt{Q}\leq\sqrt{3n}\,\tilde{\sigma}+\sqrt{10}\,LD+\sqrt{3nm}\,v_{\max}$, $L=2(n-1)mv_{\max}$ and $D^2\leq 2n$, the dominant contribution is $D\cdot LD=LD^2=O(n^2mv_{\max})$ together with $D\sqrt{n}\tilde{\sigma}=O(n\tilde{\sigma})$, giving $\mathcal{O}((n^2mv_{\max}+n\tilde{\sigma})T^{-1/4})$.

Step 7 (Nash gap). $\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^\tau)]=\frac{1}{T}\sum_t\mathbb{E}[\mathsf{NG}(\boldsymbol{\pi}^t)]\leq\frac{1}{T}\sum_t\mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)]$ by Lemma 3 (no extra factor of n). □

Remark 10 (Which set the argument runs on)**.** *Steps 1–6 use only the following about the feasible set: it is a product $\mathcal{K} = \prod_i \mathcal{K}_i$ of convex compact sets of Euclidean diameter at most D; $\bar{\Phi}$ is L-smooth and satisfies $0 \le \bar{\Phi} \le m v_{\max}$ on it; the linear oracle returns a maximizer of $\langle \mathbf{d}^t, \cdot \rangle$ over $\mathcal{K}$; and the projected estimator error obeys Lemma 7. Nothing uses that $\mathcal{K}_i$ is the full simplex $\Delta(S_i)$. In particular the argument applies verbatim on the shrunken product $\mathbf{\Delta}^\gamma$ of Corollary 2, with* FW *replaced throughout by the restricted gap* FW^γ*. The projection P_i is still the right one because $\nu_i - \tilde{\pi}_i \perp \mathbf{1}$ for all $\nu_i, \tilde{\pi}_i \in \Delta_i^\gamma$, these being differences of probability vectors.*

Optimality of the exponent within this family (Remark 7). Let $\eta = T^{-a}$ and $\rho_t \asymp t^{-p}$. Unrolling (24) gives $\|e^t\| \lesssim LD\eta/\rho_t + \tilde{\sigma}\sqrt{\rho_t} \asymp LD\eta t^p + \tilde{\sigma} t^{-p/2}$, so the middle term of (28) is $\asymp LD^2\eta T^p + D\tilde{\sigma}T^{-p/2}$. Together with the telescope term $m v_{\max} T^{a-1}$ this yields the exponent $\max\{a-1,\ p-a,\ -p/2\}$, minimized by $a = (1+p)/2$ and then by $p = 1/2$, giving $a = 3/4$ and value $-1/4$. The choice $p = 2/3$ (centralized SCG) forces $a = 5/6$ and value $-1/6$. A batched alternative, freeze $\boldsymbol{\pi}$ for $B = T^b$ rounds and average B fresh samples, gives $\max\{a+b-1,\ -a,\ -b/2\}$, again minimized at $-1/4$ ($b = 1/2$, $a = 1/4$).

Proof of Corollary 4. Set $C\,T^{-1/4} \le \varepsilon$ with $C = O(n^2 m v_{\max} + n\tilde{\sigma})$ and solve for T; substitute $\tilde{\sigma}^2 \le m v_{\max}^2$, valid because each of the $\le m$ components of (5) is a marginal contribution in $[0, v_{\max}]$. □

Proof of Lemma 8. Fix i and write $g_i = \nabla_{\pi_i}\bar{\Phi}(\tilde{\boldsymbol{\pi}})$ and $U_i = \mathrm{Unif}(S_i)$. Every $\nu_i \in \Delta_i^\gamma$ has the form $(1-\gamma)e + \gamma U_i$ for some $e \in \Delta(S_i)$, and the maximum of a linear functional over $\Delta(S_i)$ is attained at a vertex, so

$$\max_{\nu_i \in \Delta_i^\gamma} \langle g_i, \nu_i - \tilde{\pi}_i \rangle = (1-\gamma) \max_j \langle g_i, e_j \rangle + \gamma \langle g_i, U_i \rangle - \langle g_i, \tilde{\pi}_i \rangle.$$

Subtracting this from $\max_{\nu_i \in \Delta_i} \langle g_i, \nu_i - \tilde{\pi}_i \rangle = \max_j \langle g_i, e_j \rangle - \langle g_i, \tilde{\pi}_i \rangle$ gives exactly $\gamma\big(\max_j g_{i,j} - \frac{1}{|S_i|}\sum_j g_{i,j}\big) \le \gamma\big(\max_j g_{i,j} - \min_j g_{i,j}\big)$. By Lemma 2 the constant $c_i\mathbf{1}$ cancels in that difference, and the entries of $\nabla_{\pi_i} U_i$ are expected marginal contributions lying in $[0, v_{\max}]$, so the bracket is at most $v_{\max}$. Summing over i gives $\mathsf{FW}(\tilde{\boldsymbol{\pi}}) \le \mathsf{FW}^\gamma(\tilde{\boldsymbol{\pi}}) + n\gamma v_{\max}$, and $\mathsf{NG} \le \mathsf{FW}$ is Lemma 3. □

Proof of Corollary 2. The point of the proof is to run the entire argument at the *played* policy, so that no bias term arises anywhere. We first record why the naive route fails.

Step 0 (what does not work). Estimator (18) is conditionally unbiased for $\nabla_{\pi_i} U_i(\tilde{\boldsymbol{\pi}}^t)$, not for $\nabla_{\pi_i} U_i(\boldsymbol{\pi}^t)$. The naive argument would substitute its variance into Theorem 1, which is stated for iterates $\boldsymbol{\pi}^t$, and then append the comparison $\mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \le \mathsf{NG}(\boldsymbol{\pi}^t) + 4n\gamma v_{\max}$ at the end. That is invalid because the discrepancy is a bias $\nabla_{\pi_i} U_i(\tilde{\boldsymbol{\pi}}^t) - \nabla_{\pi_i} U_i(\boldsymbol{\pi}^t)$ injected into the momentum recursion at every round, which propagates through Lemma 7 and reappears in the ascent inequality of Step 2. We therefore rebuild the argument around the played policy.

Step 1 (the played sequence is itself a Frank–Wolfe trajectory). Let $\Delta_i^\gamma = \{(1-\gamma)\pi_i + \gamma\,\mathrm{Unif}(S_i) : \pi_i \in \Delta(S_i)\}$ and $\mathbf{\Delta}^\gamma = \prod_i \Delta_i^\gamma$, a product of convex compact sets of diameter $(1-\gamma)D \le D$. The played policy is $\tilde{\boldsymbol{\pi}}^t = (1-\gamma)\boldsymbol{\pi}^t + \gamma\,\mathrm{Unif} \in \mathbf{\Delta}^\gamma$. Since $\arg\max_{\nu_i \in \Delta_i^\gamma} \langle \mathbf{d}_i^t, \nu_i \rangle = (1-\gamma)e_{j_i^t} + \gamma\,\mathrm{Unif}(S_i)$ with $j_i^t = \arg\max_j [\mathbf{d}_i^t]_j$, the update of Algorithm 1 satisfies

$$\tilde{\pi}_i^{t+1} = (1-\gamma)\big[(1-\eta)\pi_i^t + \eta e_{j_i^t}\big] + \gamma\,\mathrm{Unif}(S_i) = (1-\eta)\,\tilde{\pi}_i^t + \eta\,\tilde{\mu}_i^t, \qquad \tilde{\mu}_i^t \in \arg\max_{\nu_i \in \Delta_i^\gamma} \langle \mathbf{d}_i^t, \nu_i \rangle.$$

So $\{\tilde{\boldsymbol{\pi}}^t\}$ *is* the Frank–Wolfe trajectory on $\mathbf{\Delta}^\gamma$ generated by the same $\mathbf{d}^t$; no change to the algorithm is involved, only to the set we regard it as running on.

Step 2 (no bias). For $j \in S_i$, $\mathbb{E}\big[[\widehat{\nabla}_{\pi_i}\bar{\Phi}]_j \mid \mathcal{F}_{t-1}\big] = \sum_{\mathbf{s}} \tilde{\boldsymbol{\pi}}^t(\mathbf{s})\, u_i(\mathbf{s}) \mathbf{1}\{s_i = j\}/\tilde{\pi}_i^t(j) = \mathbb{E}_{\mathbf{s}_{-i} \sim \tilde{\boldsymbol{\pi}}_{-i}^t}[u_i(j, \mathbf{s}_{-i})] = [\nabla_{\pi_i} U_i(\tilde{\boldsymbol{\pi}}^t)]_j$, and by Lemma 2 the projected estimator is unbiased for $\mathsf{P}_i \nabla_{\pi_i}\bar{\Phi}(\tilde{\boldsymbol{\pi}}^t)$, the gradient at the very iterate the recursion tracks. Lemma 7 therefore applies to $e^t := \mathsf{P}(\mathbf{d}^t - \nabla\bar{\Phi}(\tilde{\boldsymbol{\pi}}^t))$ with no bias term, and its hypothesis $\|\tilde{\boldsymbol{\pi}}^{t+1} - \tilde{\boldsymbol{\pi}}^t\| \le \eta D$ holds by Step 1.

Step 3 (variance). By Assumption 1 and $\tilde{\pi}_i^t(j) \ge \gamma/|S_i| \ge \gamma/m$,

$$\mathbb{E}\big\|\widehat{\nabla}_{\pi_i}\bar{\Phi}\big\|^2 = \sum_{j \in S_i} \frac{\mathbb{E}\big[u_i(j, \mathbf{s}_{-i})^2\big]}{\tilde{\pi}_i^t(j)} \le v_{\max}^2 \sum_{j \in S_i} \frac{1}{\tilde{\pi}_i^t(j)} \le \frac{m^2 v_{\max}^2}{\gamma},$$

so $\tilde{\sigma} \le m v_{\max}/\sqrt{\gamma}$. Note this is where forced exploration is indispensable: with $\gamma = 0$ the only available lower bound along the trajectory is $\pi_i^t(j) \ge (1-\eta)^t \pi_i^1(j)$, so $\sup_t 1/\pi_i^t(s_i)$ grows like $e^{\eta T} = e^{T^{1/4}}$ and no finite $\tilde{\sigma}$ exists.

Step 4 (apply Theorem 1 on $\mathbf{\Delta}^\gamma$). By Remark 10 the proof of Theorem 1 applies verbatim with $\mathcal{K} = \mathbf{\Delta}^\gamma$, giving $\sum_t \mathbb{E}[\mathsf{FW}^\gamma(\tilde{\boldsymbol{\pi}}^t)] = O(n^2 m v_{\max} T^{3/4} + n m v_{\max} \gamma^{-1/2} T^{3/4})$.

Step 5 (back to the unrestricted Nash gap, and optimize γ). By Lemma 8, $\mathsf{NG}(\tilde{\boldsymbol{\pi}}^t) \le \mathsf{FW}^\gamma(\tilde{\boldsymbol{\pi}}^t) + n\gamma v_{\max}$, which gives (15). Differentiating $c_1 n m v_{\max} \gamma^{-1/2} T^{3/4} + n\gamma v_{\max} T$ in γ gives $\gamma^\star = (c_1 m/2)^{2/3} T^{-1/6} = \Theta(m^{2/3} T^{-1/6})$ and value $O(n m^{2/3} v_{\max} T^{5/6})$. Requiring each of the two surviving terms of $\mathcal{R}^T/T$ to be at most ε yields (16). We retain both terms in the bound because the $T^{5/6}$ term dominates the $T^{3/4}$ term only once $T \gtrsim n^8$, so a single-term statement would be misleading at moderate horizons. □

D.9 Proof of Theorem 2

Proof. **Step 0 (Bias of $\mathbf{d}^t$).** The drift correction does not make $\mathbf{d}^t$ conditionally unbiased. Conditioning on $\mathcal{F}_{t-1}$ fixes $\mathbf{d}^{t-1}$, so (7) and Lemma 6 give

$$\mathbb{E}[\mathbf{d}^t \mid \mathcal{F}_{t-1}] = \nabla\bar{\Phi}(\boldsymbol{\pi}^t) + (1-\rho_t)\big(\mathbf{d}^{t-1} - \nabla\bar{\Phi}(\boldsymbol{\pi}^{t-1})\big), \tag{29}$$

i.e. the previous round's error is carried forward, damped by $(1-\rho_t)$. Nor is $\mathbf{d}^t$ unconditionally unbiased: iterating (29) leaves $\mathbb{E}[e^t] = \big(\prod_{s\le t}(1-\rho_s)\big)\mathbb{E}[e^0]$, which is nonzero because $\mathbf{d}^0 = \mathbf{0}$. What the drift correction actually buys is that the *increment* injected at round t is conditionally mean-zero, which is all the mean-square recursion needs. We therefore work with the error recursion directly and never invoke unbiasedness of $\mathbf{d}^t$.

Step 1 (Filtration and sampling construction). Let $\mathcal{F}_{t-1} = \sigma\big(\{\mathbf{s}^r, \alpha^r, \{c_k^r\}, \{a_k^r, b_k^r\}, \text{oracle noise at round } r\}_{r<t}\big)$, so that $\boldsymbol{\pi}^{t-1}, \boldsymbol{\pi}^t, \boldsymbol{\mu}^{t-1}$ and $\mathbf{d}^{t-1}$ are all $\mathcal{F}_{t-1}$-measurable ($\boldsymbol{\pi}^t$ is computed from $\mathbf{d}^{t-1}$ at the end of round $t-1$; in particular b_k^t, the task supported by μ_k^{t-1}, is $\mathcal{F}_{t-1}$-measurable and *not* random at round t). At round t the following are drawn, mutually independently and independently of $\mathcal{F}_{t-1}$:

(S1) the announcements $a_k^t \sim \pi_k^{t-1}$, one per agent k, each generated by agent k;

(S2) $\mathbf{s}^t \sim \boldsymbol{\pi}^t$, the played profile, used only for the gradient estimator (5);

(S3) agent i's own $\alpha^t \sim \mathrm{Unif}[0,1]$ and coins $c_k^t \sim \mathrm{Bern}(\alpha^t \eta)$, which together with (S1) determine $\mathbf{s}'^t$ by (10) and are used only for the drift correction (11);

(S4) one fresh oracle-noise variable per *call*, independent across calls, so that the gradient-query and drift-query streams are independent even where they touch the same (task, coalition) pair.

Two points deserve emphasis. First, nobody samples a distribution they do not own: the randomness carrying π_k^{t-1} into the estimator is generated by agent k in (S1), and agent i contributes only (S3). Lemma 6, Step 1, shows the assembled $\mathbf{s}'^t$ nonetheless has law $\boldsymbol{\pi}^t(\alpha^t)$. Second, independence of the profile draws is *not* by itself enough to kill the full cross term in Step 3, the estimators can be coupled through shared oracle noise, which is why (S4) is stated per call and not per distinct pair; Appendix B.3 also records what the bound becomes if common random numbers are used instead (the exponent is unchanged).

Step 2 (Error recursion). Let $e^t = \mathsf{P}(\mathbf{d}^t - \nabla\bar{\Phi}(\boldsymbol{\pi}^t))$,

$$\zeta^t := \mathsf{P}\big(\tilde{\mathbf{\Delta}}^t - (\nabla\bar{\Phi}(\boldsymbol{\pi}^t) - \nabla\bar{\Phi}(\boldsymbol{\pi}^{t-1}))\big), \qquad \xi^t := \widehat{\nabla} U(\boldsymbol{\pi}^t) - \nabla U(\boldsymbol{\pi}^t).$$

Substituting (7) and splitting $-\mathsf{P}\nabla\bar{\Phi}(\boldsymbol{\pi}^t)$ as $-(1-\rho_t)\mathsf{P}\nabla\bar{\Phi}(\boldsymbol{\pi}^t) - \rho_t \mathsf{P}\nabla\bar{\Phi}(\boldsymbol{\pi}^t)$ gives the *exact* identity

$$e^t = (1-\rho_t)\big(e^{t-1} + \zeta^t\big) + \rho_t\, \mathsf{P}\xi^t. \tag{30}$$

Step 3 (Mean square). By Lemma 6 and (S2), $\mathbb{E}[\zeta^t \mid \mathcal{F}_{t-1}] = 0$; by (S1) and Lemma 5, $\mathbb{E}[\mathsf{P}\xi^t \mid \mathcal{F}_{t-1}] = 0$. Expanding (30),

$$\mathbb{E}\|e^t\|^2 = (1-\rho_t)^2\Big(\mathbb{E}\|e^{t-1}\|^2 + 2\,\mathbb{E}\langle e^{t-1}, \zeta^t\rangle + \mathbb{E}\|\zeta^t\|^2\Big) + 2\rho_t(1-\rho_t)\,\mathbb{E}\langle e^{t-1} + \zeta^t,\, \mathsf{P}\xi^t\rangle + \rho_t^2\,\mathbb{E}\|\mathsf{P}\xi^t\|^2.$$

The first cross term vanishes because e^{t-1} is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[\zeta^t \mid \mathcal{F}_{t-1}] = 0$ (Lemma 6). The second vanishes by the tower property applied to $\mathcal{G}_t := \sigma(\mathcal{F}_{t-1}, \alpha^t, \{c_k^t\}, \{a_k^t\}, \text{drift-query noise})$: conditionally on $\mathcal{G}_t$, $e^{t-1} + \zeta^t$ is fixed while $\mathsf{P}\xi^t$ still has mean zero, because (S2) and the gradient-query noise in (S4) are independent of $\mathcal{G}_t$. *Both ingredients are needed here: the independence of the play draw from the drift draw, and the per-call independence of the oracle noise. Independent draws alone leave ζ^t and ξ^t coupled through shared noise; Appendix B.3 bounds that residual covariance by $n\eta\sigma_H\tilde{\sigma}$ and shows it is absorbed into Q_B without changing the exponent.* Hence, using $\mathbb{E}\|\zeta^t\|^2 = \sum_i \mathbb{E}\|\zeta_i^t\|^2 \le n\eta^2\sigma_H^2$ by Assumption 1, the factor n is the aggregation over the n agents, together with $(1-\rho_t)^2 \le 1-\rho_t$,

$$\mathbb{E}\|e^t\|^2 \le (1-\rho_t)\,\mathbb{E}\|e^{t-1}\|^2 + n\sigma_H^2\eta^2 + \rho_t^2\, n\tilde{\sigma}^2. \tag{31}$$

The essential difference from Variant A is that the middle term is $O(\eta^2)$ rather than $O(\eta^2/\rho_t)$: the drift correction removes the momentum *lag*, leaving only its *noise*. Note that (31) is obtained without any claim of unbiasedness for $\mathbf{d}^t$, the residual term in (29) is exactly the $(1-\rho_t)e^{t-1}$ that the recursion already tracks.

Step 4 (Induction). With $\rho_t = (t+1)^{-2/3}$, induction gives $\mathbb{E}\|e^t\|^2 \le Q_B/(t+2)^{2/3}$ for $Q_B = 3n\tilde{\sigma}^2 + 6n\sigma_H^2 + 2nmv_{\max}^2$ and $\eta = T^{-2/3}$. Assuming the bound at $t-1$,

$$\mathbb{E}\|e^t\|^2 \le \frac{Q_B}{(t+1)^{2/3}} - \frac{Q_B}{(t+1)^{4/3}} + n\sigma_H^2\eta^2 + \frac{n\tilde{\sigma}^2}{(t+1)^{4/3}}.$$

Since $(t+1)^{-2/3} - (t+2)^{-2/3} \le \frac{2}{3}(t+1)^{-5/3} \le 0.53\,(t+1)^{-4/3}$, it suffices that $n\sigma_H^2\eta^2(t+1)^{4/3} + n\tilde{\sigma}^2 \le 0.47\,Q_B$. For $t \le T$, $(t+1)^{4/3} \le (2T)^{4/3} \le 2.52\,T^{4/3}$, and $\eta^2 T^{4/3} = 1$, so the left side is at most $2.52\,n\sigma_H^2 + n\tilde{\sigma}^2$, and $0.47\,Q_B = 1.41n\tilde{\sigma}^2 + 2.82n\sigma_H^2 + 0.94nmv_{\max}^2$ dominates it termwise. The base case $\mathbb{E}\|e^0\|^2 \le nmv_{\max}^2 \le Q_B/2^{2/3}$ holds. Note $\eta = T^{-2/3}$ is exactly what makes Q_B free of T given $\rho_t \asymp t^{-2/3}$; the pairing $(\eta, \rho_t) = (T^{-1/2}, t^{-1/2})$ leaves $\sigma_H^2 D^2\sqrt{t}/T$ in $\mathbb{E}\|e^t\|^2$ and degrades the final rate to $T^{-1/4}$.

Step 5 (FW ascent). Steps 1–4 of the proof of Theorem 1 apply verbatim, they use only smoothness, boundedness of $\bar{\Phi}$ and the linear oracle, none of which depend on the estimator. With $\sum_{t=1}^{T}(t+2)^{-1/3} \le \frac{3}{2}(T+2)^{2/3} \le 2.4\,T^{2/3}$,

$$\frac{1}{T}\sum_t \mathbb{E}[\mathsf{FW}(\boldsymbol{\pi}^t)] \le \frac{mv_{\max}}{\eta T} + \frac{4.8\,D\sqrt{Q_B}}{T^{1/3}} + \frac{L\eta D^2}{2},$$

which with $\eta = T^{-2/3}$ is (14). Substituting $\sqrt{Q_B} \le \sqrt{3n}\tilde{\sigma} + \sqrt{6n}\,\sigma_H + \sqrt{2nm}\,v_{\max}$ and $D \le \sqrt{2n}$ gives $D\sqrt{Q_B} \le \sqrt{6}\,n\tilde{\sigma} + 2\sqrt{3}\,n\sigma_H + 2n\sqrt{m}\,v_{\max}$, and since $\sigma_H \ge \sqrt{m}v_{\max}$ whenever $n \ge 2$ the last term is absorbed, giving $\mathcal{O}((mv_{\max} + n\tilde{\sigma} + n\sigma_H)T^{-1/3})$; the $L\eta D^2$ term is $O(n^2 mv_{\max}T^{-2/3})$ and lower order.

Step 6 (Optimality of the exponent). Repeating the calculation of the paragraph after Theorem 1 with the $O(\eta^2)$ drift noise, $\eta = T^{-a}$, $\rho_t \asymp t^{-p}$, gives exponent $\max\{a-1,\ p/2-a,\ -p/2\}$, minimized at $p = a = 2/3$ with value $-1/3$. This recovers exactly the schedule and exponent of Theorem 2 of Zhang et al. (2020), as it should: that theorem is also stated for a smooth nonconvex objective under a constant step size with a randomly selected output iterate, and we are not improving on it. The $\mathcal{O}(\sqrt{T})$ rate in the same paper belongs to its *convex* branch, which uses a decaying step size $\eta_t \asymp 1/t$ and a contraction on the optimality gap; neither is available here, the first because Nash regret sums the iterates uniformly, the second because $\bar{\Phi}$ is never concave on $\boldsymbol{\Delta}$ (Proposition 1). □

Proof of Corollary 3. Set (14) $\le \varepsilon$, substitute $\tilde{\sigma}^2 \le mv_{\max}^2$ and $\sigma_H^2 \le 4(n-1)^2 mv_{\max}^2$ (Proposition 2), and solve. The dominant term is $n\sigma_H = O(n^2\sqrt{m}\,v_{\max})$, giving $T = O(n^6 m^{3/2} v_{\max}^3/\varepsilon^3)$. □

Proof of Proposition 2. Fix $i, j \in S_i$ and $k \ne i$. The summand $\Phi(j, b_k^t, \mathbf{s}'^t_{-\{i,k\}}) - \Phi(j, a_k^t, \mathbf{s}'^t_{-\{i,k\}})$ is a difference of potentials at two profiles that agree outside coordinate k. Moving one agent from task a_k^t to task b_k^t alters exactly two coalitions, and each task value lies in $[0, v_{\max}]$ almost surely

by Assumption 1, so the summand lies in $[-2v_{\max}, 2v_{\max}]$. Summing the $n-1$ summands and multiplying by η,

$$\left|[\tilde{\boldsymbol{\Delta}}_i^t]_j\right| \le 2\eta(n-1)v_{\max} \quad \text{for every } j \in S_i,$$

$$\text{hence} \qquad \|\tilde{\boldsymbol{\Delta}}_i^t\| \le 2\eta(n-1)\sqrt{|S_i|}\, v_{\max} \le 2\eta(n-1)\sqrt{m}\, v_{\max}.$$

Since P_i is an orthogonal projection and $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$, the quantity in (12) is at most $4\eta^2(n-1)^2 m v_{\max}^2$, i.e. $\sigma_H^2 \le 4(n-1)^2 m v_{\max}^2$. No property of v_j beyond boundedness is used, so the bound is insensitive to how identity-dependent the coalition values are. The same count gives the per-round cost: $|S_i|(n-1) \le m(n-1)$ summands, each two marginal contributions, i.e. $O(mn)$ oracle calls and $O(mn)$ arithmetic, with $O(m)$ state carried between rounds. Note this bound is deliberately crude: it treats all $n-1$ contributions as simultaneously extremal and perfectly aligned. Once the Frank–Wolfe directions stabilize, $a_k^t = b_k^t$ for most k and the corresponding summands are identically zero, which is part of why the observed convergence is far faster than certified (Remark 4). □

D.10 Proof of Corollary 5

Proof. Under anonymity only $m(n+1)$ marginal contributions exist, so the feature map $\phi_i(\cdot)$ of Section A.8 is well defined with $d = m(n+1)$ and $\nabla_{\pi_i}\bar{\Phi}(\boldsymbol{\pi})(j) = \langle \phi_i(j), \theta(\boldsymbol{\pi}_{-i})\rangle + c_i$. Assumption 2 is exactly the hypothesis that the windowed estimator (21) tracks the moving target well enough that $\mathbb{E}\|\hat{\theta}^t - \theta(\boldsymbol{\pi}_{-i}^t)\|^2 \le C\,\mathrm{poly}(n,m)v_{\max}^2\varrho_t$ with $\frac{1}{T}\sum_t \sqrt{\varrho_t}$ of the stated order. Given it, the induced gradient and drift errors satisfy $\tilde{\sigma}^2, \sigma_H^2 = \mathcal{O}(\mathrm{poly}(n,m)v_{\max}^2)$ and the error-sum step (Step 5 of the proof of Theorem 1, Step 5 of the proof of Theorem 2) goes through with those constants, giving (22) and (23). The drift correction uses $\hat{\theta}^t - \hat{\theta}^{t-1}$ and therefore needs neither auxiliary announcements nor sampled potential differences. The exponents of T are unchanged; what improves relative to the general case are the $\mathrm{poly}(n,m)$ constants and the fact that a model of v_j is representable at all.

This result is conditional on Assumption 2. We specify the estimator in (21), quantify the target's per-round motion as $O(\eta\,\mathrm{poly}(n,m)v_{\max})$, and give the window $W = \Theta(\eta^{-2/3})$ that balances tracking bias against estimation variance. A proof of Assumption 2 for (21) under Frank–Wolfe dynamics remains open. The unconditional results of this paper are Theorems 1 and 2 and Corollaries 4–3. □

E Additional experimental results

Every result here uses the protocol of Section 4: exact metric evaluation at every round on the played policy, resampled instances crossed with algorithm seeds, and the schedules of Algorithm 1. Ablations use 9 runs. All of it was regenerated with the revised Variant B (sampled drift correction); no number here is inherited from the earlier implementation. Each table shows $\mathcal{R}^T/T$, the quantity the theorems bound, alongside the final-iterate gap.

Momentum-error decay. Figure 2 plots the projected momentum error $\mathbb{E}\|e^t\|^2 = \mathbb{E}\|\mathsf{P}(\mathbf{d}^t - \nabla\bar{\Phi}(\boldsymbol{\pi}^t))\|^2$. This is the exact quantity Lemma 7 and Step 3 of Theorem 2 bound, computed against the exact gradient. Both estimators decay monotonically and the drift-corrected Variant B sits below Variant A throughout, which is the qualitative content of the two recursions. Both decay *faster* than certified: the fitted log-log slopes over $t \ge 50$ are -1.13 (Variant A, certified $t^{-1/2}$) and -1.12 (Variant B, certified $t^{-2/3}$), because $\|\boldsymbol{\pi}^t - \boldsymbol{\pi}^{t-1}\|$ falls far below the worst-case ηD once the linear oracle stabilizes.

Figure 1a nevertheless shows close agreement with the predicted average-gap exponents. The bound of (28) is a sum of three terms, and the one that binds is the telescope $mv_{\max}/(\eta T)$, which depends only on the step size and the range of $\bar{\Phi}$, not on the estimator at all. With a constant $\eta = T^{-3/4}$ the iterate simply cannot traverse the simplex faster than ηT allows, so the average gap decays at $T^{-1/4}$ however good the gradient estimates become. This is why the certified exponents are observed even though the momentum error beats its own bound, and it identifies the step size as the term that sets the rate; estimator noise does not. This is also consistent with the last iterate converging far faster than the average (Section 4).

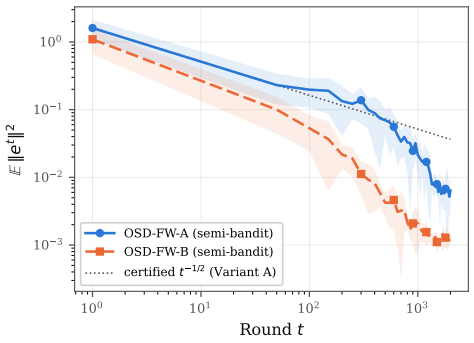

Figure 2: Momentum-error decay, identity-dependent FVG, n=8, m=5, T=2000. Shaded: ± 1 s.d. over 9 runs. Dotted: the exponent certified for Variant A.

Exploration weight γ (Variant A, pure bandit). Table 8 sweeps γ at T=1500. The response has the shape Corollary 2 predicts, a $\gamma^{-1/2}$ variance term punishing small γ and an $n\gamma v_{\max} T$ oracle-restriction term punishing large γ, but the empirical optimum over the grid is at its lower end, $\gamma = 0.01$ on $\mathcal{R}^T/T$ (0.1476) and $\gamma = 0.05$ on the final gap (0.1042), with the two within 4% of each other, well below $\gamma^\star = (m/4)^{2/3} T^{-1/6} = 0.343$, which over-explores by roughly an order of magnitude and costs about 30% in $\mathcal{R}^T/T$ and 50% in final gap. The worst-case constant in $\gamma^\star$ is conservative: the $n\gamma v_{\max}$ term of Lemma 8 charges the full gradient range at every round, which is loose whenever the equilibrium is not near a simplex face. Practitioners should tune γ; the theoretical value is what makes the $T^{5/6}$ guarantee hold uniformly over instances. Note this ablation also bounds how much of the pure-bandit variant's weak showing in Table 4 is attributable to over-exploration. Tuning γ narrows the gap but does not close it, so the feedback model also contributes.

Table 8: Exploration ablation, `OSD-FW-A` (bandit), identity-dependent FVG, n=8, m=5, T=1500, 9 runs per row. $\mathcal{R}^T/T$ is the quantity Corollary 2 bounds.

γ	$\mathcal{R}^T/T \downarrow$	Final Nash gap $\downarrow$	Cumulative Nash regret $\downarrow$
0.01	0.1476 ± 0.0164	0.1055 ± 0.0146	221.4 ± 24.6
0.05	0.1522 ± 0.0192	0.1042 ± 0.0074	228.4 ± 28.8
0.1	0.1577 ± 0.0200	0.1118 ± 0.0073	236.5 ± 30.1
0.2	0.1725 ± 0.0217	0.1242 ± 0.0120	258.8 ± 32.5
0.4	0.2045 ± 0.0233	0.1671 ± 0.0232	306.7 ± 34.9
$\gamma^\star = 0.343$	0.1958 ± 0.0238	0.1556 ± 0.0229	293.8 ± 35.7

Scalability in n. Table 9 scales n from 4 to 12 on the anonymous FVG at T=1500 (the exact-*evaluation* cost is $\Theta(nm2^n)$, which caps the range at which we can report exact Nash gaps; the algorithm itself is $\mathrm{poly}(n, m)$ per round). The final gap grows mildly while $\mathcal{R}^T/T$ *decreases* monotonically, from 0.065 at n=4 to 0.035 at n=12, because with more agents on the same m=5 tasks the equilibrium is less sensitive to any single agent's policy. Per-run wall-clock grows from 0.44s to 3.05s over the range, a factor of 6.9 for a factor of 3 in n; this is dominated by the $\Theta(nm2^n)$ measurement, not by the $O(mn)$-per-round algorithm, and should not be read as the algorithm's scaling.

Feasibility-structured environments. Figure 3 presents the crowdsourcing-style and MRTA-style environments, in which both the feasible sets S_i and the identity weights W_{ij} are induced by the same underlying geometry (spatial proximity; capability–requirement coverage). Here `OSD-FW-B` is best on $\mathcal{R}^T/T$ on both: 0.0260 (crowdsourcing) and 0.0228 (MRTA) against 0.0301 and 0.0347 for the next best, UCB-MC. Its final gap is below 10^{-4} on both, though on the crowdsourcing instance UCB-MC also reaches that floor. These environments test heterogeneous, overlapping feasible sets in addition to identity-dependent values; they are synthetic, and we do not present them as replications of the cited datasets.

Table 9: Scalability of OSD−FW-A (unilateral-oracle), anonymous FVG, m=5, T=1500, 9 runs per row. Wall-clock is per run on one CPU core and includes the exact-evaluation cost, which is $\Theta(nm2^n)$ and is a property of the *measurement*, not of the algorithm.

n	$\mathcal{R}^T/T \downarrow$	Final Nash gap $\downarrow$	Cumulative Nash regret $\downarrow$	Wall-clock (s)
4	0.0654 ± 0.0026	0.0115 ± 0.0056	98.2 ± 3.9	0.44
6	0.0541 ± 0.0021	0.0288 ± 0.0098	81.1 ± 3.2	0.77
8	0.0451 ± 0.0009	0.0356 ± 0.0093	67.6 ± 1.4	1.22
10	0.0386 ± 0.0024	0.0332 ± 0.0079	57.8 ± 3.6	1.84
12	0.0347 ± 0.0016	0.0318 ± 0.0088	52.1 ± 2.3	3.05

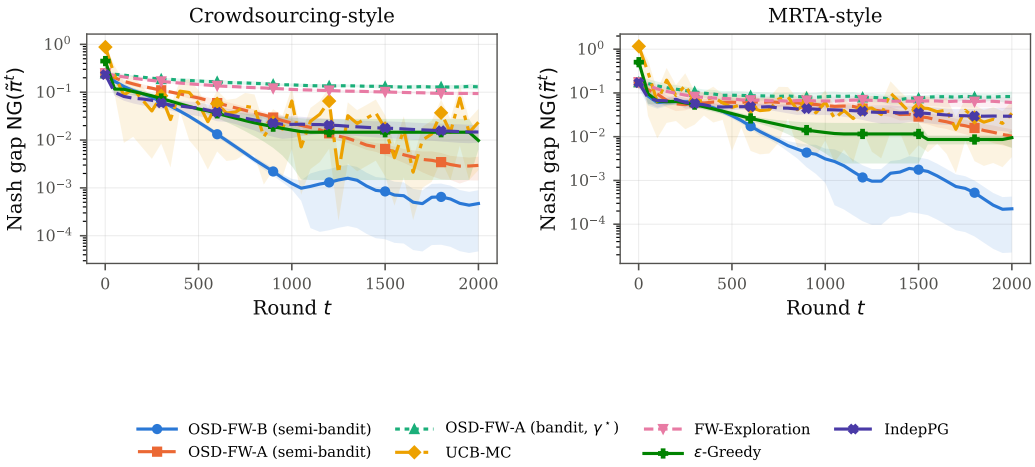

Figure 3: Feasibility-structured synthetic environments, n=8, m=5, T=2000, 9 runs. Shaded: ± 1 s.d.; y log-scaled.

Outside the assumptions: a non-monotone environment. The value function $v_j(C) = (\sum_{i\in C} W_{ij})/\sqrt{|C|}$ is *not* monotone: adding an agent whose weight is below $(\sum_{k\in C} W_{kj})(\sqrt{1+1/|C|}-1)$ strictly lowers the coalition's value, so this environment violates Definition 1. We include it deliberately, because the theory predicts what should happen. Lemma 1 and Corollary 1 use only the definition of marginal contributions and therefore still hold: the game is still an exact potential game and Frank–Wolfe stationary points are still Nash equilibria. Only the explicit smoothness constant $L = 2(n-1)mv_{\max}$ and part (c) of Proposition 1 use submodularity. Table 10 bears this out: both `OSD-FW` variants converge (final gaps 0.0050 ± 0.0140 and 0.0033 ± 0.0028), while FW-Exploration, whose analysis leans on the congestion-style structure, stalls at 0.0923. On $\mathcal{R}^T/T$ `OSD-FW`-B leads at 0.0450 against UCB-MC's 0.0544. The large deviation on `OSD-FW`-B's final gap comes from a minority of runs that settle at a different equilibrium; with monotonicity gone, the potential has stationary points absent from monotone instances. We therefore interpret these runs as convergence to different equilibria.

Table 10: Non-monotone environment $v_j(C) = (\sum_{i\in C} W_{ij})/\sqrt{|C|}$, n=8, m=5, T=2000, 9 runs. This instance lies *outside* the paper's assumptions.

Algorithm	$\mathcal{R}^T/T \downarrow$	Final Nash gap $\downarrow$	Cumulative Nash regret $\downarrow$
`OSD-FW`-B	0.0450 ± 0.0117	0.0050 ± 0.0140	90.0 ± 23.4
`OSD-FW`-A (unilateral)	0.0815 ± 0.0095	0.0033 ± 0.0028	163.0 ± 19.0
UCB-MC	0.0544 ± 0.0104	0.0517 ± 0.1302	108.8 ± 20.8
FW-Exploration	0.1014 ± 0.0092	0.0923 ± 0.0079	202.7 ± 18.4

Corrected versus original schedules. The analysis pins particular step-size and momentum exponents, so it is fair to ask whether they are the empirically best choice. Table 11 shows that they are not. On the identity-dependent instance the alternative pairings beat the ones we can prove bounds for: $\eta{=}T^{-2/3}$, $\rho_t{\asymp}t^{-2/3}$ gives Variant A an $\mathcal{R}^T/T$ of 0.0356 against 0.0636 for the paper's schedule, and $\eta{=}T^{-1/2}$, $\rho_t{\asymp}t^{-1/2}$ gives Variant B 0.0109 against 0.0328, a factor of three in both cases. The schedules in Algorithm 1 are the ones for which the argument closes, under the centralized-SCG pairing the constant Q in Lemma 7 carries a factor $T^{2/3}$ and Theorem 1 degrades to $T^{-1/6}$, but a larger step size evidently pays on benign instances, where the telescope term that drives the certified rate is the binding constraint. Finding schedules that are both provable and empirically competitive is a concrete open direction.

Table 11: Schedule comparison, identity-dependent FVG, n=8, m=5, T=2000, 9 runs per row.

Variant, (η, ρ_t)	$\mathcal{R}^T/T \downarrow$	Final Nash gap $\downarrow$	Cumul. regret $\downarrow$
A, paper ($T^{-3/4}$, $t^{-1/2}$)	0.0636 ± 0.0081	0.0033 ± 0.0025	127.2 ± 16.1
A, centralized-SCG ($T^{-2/3}$, $t^{-2/3}$)	0.0356 ± 0.0039	0.0004 ± 0.0007	71.3 ± 7.7
B, paper ($T^{-2/3}$, $t^{-2/3}$)	0.0328 ± 0.0041	0.0000 ± 0.0000	65.7 ± 8.2
B, alternative ($T^{-1/2}$, $t^{-1/2}$)	0.0109 ± 0.0013	0.0000 ± 0.0000	21.8 ± 2.6

Reproducibility. The released code contains `osd_fw.py` (game and `OSD-FW`), `exact_eval.py` (exact mask-summation evaluation), `baselines.py` (the comparison methods, including a transcription of Dong et al. (2024), Algorithm 1), `runner.py`, `runner_appendix.py` and `runner_budget.py` (checkpointed experiment drivers writing one JSON record per run), `aggregate.py`, `aggregate_appendix.py`, `aggregate_budget.py` and `make_tables_figs.py` (which emit all reported tables and figures directly from those records), and `verify_theory.py` (numerical checks of Lemma 6 and Lemma 8 against brute-force enumeration). Every number reported here is produced by that pipeline from a fixed set of instance and algorithm seeds; there is a single authoritative result set and no legacy files.