The Deceptive Bandit Problem: Exploratory Coupling and the Fragility of Multi-Agent Learning

Michael Tang, Mahmoud Abdelgalil, Jorge I. Poveda

***Abstract*— Randomized exploration is central to bandit learning, multi-agent reinforcement learning, and zeroth-order policy search, yet its independence and privacy are usually only treated as technical assumptions. We show that these properties are critical for security purposes and demonstrate how an adversarial agent can exploit privileged information on another agent's exploration. We analyze a deceiver-victim pair in the minimal two-player strongly monotone setting, where a deceptive player obtains leaked signals that are merely *correlated* with the victim's exploration. We show that, by coupling their own exploratory action with this information, the deceptive player injects an externality that steers the learning dynamics to a new steady state, called the *deceptive Nash equilibrium (DNE)*. We prove that the deceptive bandit learning (DBL) dynamics converge to an arbitrarily small neighborhood of the DNE while retaining optimal convergence rates. Interestingly, our analysis attains these optimal rates while relaxing second-order smoothness conditions from standard bandit optimization literature. We characterize conditions under which deception strictly shifts the steady state and its effect on the deceiver's cost, illustrating the results in a resource-allocation game.**

I. Introduction

Multi-agent learning has attracted significant attention in recent years with applications spanning many domains. In many of these applications, agents do not have access to gradient information and must instead rely solely on *bandit feedback* [1], i.e., measurements of their cost/payoff functions, to update their policies. This has motivated a rich body of literature on *multi-agent bandit learning*, where agents simultaneously explore their action spaces and exploit observed costs/payoffs to find optimal policies. Example applications include resource allocation [2], distributed and networked systems [3]–[5], robotics [6], and financial decision-making [7], [8]. More broadly, randomized exploration is a basic ingredient of many learning methods, including stochastic approximation [9], multi-agent reinforcement learning [10], and zeroth-order policy search [11]. If agents learn through bandit feedback, the convergence to the Nash equilibrium (NE) [12]–[14] relies on the assumption that the exploratory policies used by different agents are *independent* and *private*. However, in an open, networked, or competitive environment, such an assumption is fragile; an adversarial agent can *intercept* or *infer* privileged information about the exploratory actions of other agents [15]–[17].

M. Tang and J. I. Poveda are with the Electrical and Computer Engineering Department at the University of California, San Diego, La Jolla, CA, USA. M. Abdelgalil is with the Mechanical and Aerospace Engineering Department at the University at Buffalo, State University of New York, Buffalo, NY, USA.

This work was supported in part by NSF grants DGE-2545911, CMMI 2228791, and the DARPA grant HR0011-25-3-0225.

Contributions: We expose a security vulnerability in exploration-driven multi-agent learning: information correlated with an agent's exploratory behavior can be amplified into a persistent bias that changes the long-term outcome of learning. To isolate the core mechanism, we formalize it in two-player strongly monotone games, where a deceptive player modifies only its own exploration using a noisy signal correlated with the oblivious player's current dither; extensions to N-player settings are also discussed. This unilateral coupling induces an externality in the oblivious player's gradient estimate, scaled by the dither-signal correlation term. We prove that the resulting *deceptive bandit learning* (DBL) dynamics converge to a small neighborhood of a new steady state, a *deceptive* Nash equilibrium (DNE), with non-asymptotic mean-square guarantees. Our analysis attains the optimal convergence rates from [18] while relaxing their second-order smoothness conditions on the players' costs; to our knowledge, this relaxation is also novel in the deception-free setting. We characterize when deception strictly shifts the equilibrium and provide geometric conditions under which this shift improves the deceiver's steady-state cost. Conversely, we also outline some defense strategies against deception that eliminate the induced bias. We then illustrate our main results on a resource-allocation game.

Related Literature: Other works that study the vulnerabilities of multi-agent learning algorithms include utility measurement poisoning [19], [20], adversarial policies [21], and continuous-time equilibrium seeking [22], [23]. Unlike [19]–[21], the deceiver in our setting never tampers with the victim's cost measurements, but manipulates learning solely through its own feasible exploratory action. In [22], the deterministic sinusoidal probe structure means that knowledge of the victim's exploration frequency suffices to reproduce its exploration signal; [23] requires exact real-time measurements of an OU-driven probe. Both study unconstrained continuous-time dynamics and establish local practical asymptotic guarantees under multiple vanishing parameter limits. In contrast, our setting does not require exact reconstruction or real-time measurements of the exploration signal; mere correlation suffices, as our deceiver receives only a noisy, possibly intermittent, randomly transformed, history-adaptive signal correlated with the current dither. We show that the conditional cross-covariance between leakage and exploration exactly governs the induced bias in the victim's gradient estimator. For constrained, discrete-time bandit play, this yields, to our knowledge, the first non-asymptotic analysis of adversarial exploration coupling in multi-agent learning, with mean-square convergence guarantees.

II. Preliminaries

Given a matrix $A \in \mathbb{R}^{n\times m}$, we use $\|A\|$ to denote the operator norm of A. If A is a vector, then $\|A\|$ evaluates to the Euclidean norm of A. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, then $\nabla f : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ denotes the transpose of the Jacobian of f. We use $\mathbb{S}^n = \{s \in \mathbb{R}^n : \|s\| = 1\}$ and $\mathbb{B}^n = \{s \in \mathbb{R}^n : \|s\| \le 1\}$ to denote the unit sphere and ball in $\mathbb{R}^n$, respectively. We denote the ball around $p \in \mathbb{R}^n$ of radius $r > 0$ by $\mathbb{B}_r(p) = \{s \in \mathbb{R}^n : \|p - s\| \le r\}$. For a closed convex set $E \subset \mathbb{R}^n$, the projection operator is defined as $\text{Proj}_E(x) = \arg\min_{p\in E} \|x - p\|$. We use int E to denote the interior of the set E.

III. Problem Setup

We focus on a two-player game, the minimal setting containing a deceiver-victim pair; the generalization to N-player settings is discussed in Remark 2. Player i selects an action x_i from an action space $\mathcal{X}_i \subset \mathbb{R}^n$ with the goal of minimizing their cost $J_i : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$. The joint action profile is given by $x = (x_1, x_2) \in \mathcal{X}$, where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ denotes the game's joint action space. A point x^* is said to be a *Nash equilibrium (NE)* if each player has no incentive to unilaterally change their action, i.e.,

$$x_i^* \in \arg\min_{x_i\in\mathcal{X}_i} J_i(x_i; x_{3-i}^*), \quad i \in \{1, 2\}, \tag{1}$$

where we use $J_i(x_i; x_{3-i})$ and $J_i(x)$ interchangeably. Given the costs, the *pseudogradient* of the game is defined as:

$$\mathcal{G}(x) := \begin{bmatrix} \nabla_1 J_1(x) \\ \nabla_2 J_2(x) \end{bmatrix}, \tag{2}$$

where $\nabla_i(\cdot)$ represents differentiation with respect to x_i. We impose the following assumptions on the game parameters:

Assumption 1. *For each $i \in \{1, 2\}$ and $j \in \{1, 2\}$, the following hold:*

1) *The set $\mathcal{X}_i$ is compact and convex, and there exists $p_i \in \mathcal{X}_i$, $r_i > 0$ such that $\mathbb{B}_{r_1}(p_1) \times \mathbb{B}_{r_2}(p_2) \subset \mathcal{X}$.*
2) *The function $J_i(x)$ is continuously differentiable on an open neighborhood of $\mathcal{X}$, and there exists $L_{ij} > 0$ such that the function $\nabla_j J_i(x)$ is L_{ij}-Lipschitz on $\mathcal{X}$.*
3) *There exists $\gamma > 0$ such that the operator $\mathcal{G}(\cdot)$ is γ-strongly monotone over $\mathcal{X}$, i.e., the following holds*

$$\langle \mathcal{G}(x) - \mathcal{G}(y),\, x - y\rangle \ge \gamma \|x - y\|^2, \tag{3}$$

for all $x, y \in \mathcal{X}$. □

Conditions 1-3 are standard in continuous-action bandit learning [1], [24], [25]; note that only first-order smoothness is imposed. Games with pseudogradients that satisfy (3) are referred to as γ-*strongly monotone*, which are ubiquitous in the game theory literature [24], [26], [27]. It is well-known that in games where Assumption 1 holds, the Nash equilibrium x^* is unique and satisfies the following variational inequality

$$\langle \mathcal{G}(x^*),\, x - x^*\rangle \ge 0, \quad \forall x \in \mathcal{X}. \tag{4}$$

Hence, we say that an action profile $x^* \in \mathcal{X}$ is a "*Nash equilibrium of* $(\mathcal{G}, \mathcal{X})$" if (4) holds.

A. Standard Bandit Feedback Algorithm

For our work, we will study the bandit feedback-based algorithm from [1], [18], [25], which can be implemented by the agents to learn the Nash equilibrium by only using measurements of their cost functions. At each time step k, player i samples a vector $\mu_{i,k} \in \mathbb{S}^n$ uniformly at random and computes the following intermediate exploration step

$$\bar{x}_{i,k} = x_{i,k} + \delta\mu_{i,k}, \tag{5}$$

where $\delta \in (0, \min_i r_i)$ is the search radius and $x_{i,k}$ is the action of player i at the k-th iteration. For notational convenience, we assume all players use the same perturbation radius δ. This is quite standard [18], [25], and the case of heterogeneous δ_i is discussed in Remark 2. Player i then performs the action $\bar{x}_{i,k}$, observes their resulting cost $J_i(\bar{x}_k)$ (with $\bar{x}_k = (\bar{x}_{1,k}, \bar{x}_{2,k})$), and updates their action:

$$x_{i,k+1} = \text{Proj}_{\mathcal{X}_i^\delta}\left(x_{i,k} - \eta_k \frac{n}{\delta} J_i(\bar{x}_k)\mu_{i,k}\right), \tag{6}$$

where $\mathcal{X}_i^\delta = p_i + (1 - \frac{\delta}{r_i})(\mathcal{X}_i - p_i)$; intuitively, this is a homothetic contraction of $\mathcal{X}_i$ towards the point p_i (from Assumption 1). The projection operator in (6) ensures that the action profile at the next iterate, specified in (5), remains feasible. This type of feasibility adjustment, which is common in bandit learning [25], generalizes the contraction $(1 - \delta)\mathcal{X}_i$ from [18] and allows us to consider action spaces that need not contain the origin. Both players use a common learning rate η_k that satisfies the standard assumption [9]:

Assumption 2. *The sequence $\{\eta_k\}_{k=1}^\infty$ satisfies:*

$$\lim_{k\to\infty} \eta_k = 0, \quad \sum_{k=1}^{\infty} \eta_k = \infty, \tag{7}$$

and $\eta_k > 0$ for each $k \in \mathbb{N}$. □

The algorithm (5)-(6) has been covered extensively in the *bandit optimization* literature [1], [18], [25], [26], and the main intuition behind it is in the signal $\frac{n}{\delta} J_i(\bar{x}_k)\mu_{i,k}$ from (6), which serves as a single-point stochastic gradient estimator. More precisely, its conditional expectation equals the gradient of a suitably smoothed version of J_i, which differs from $\nabla_i J_i(x_k)$ by an $\mathcal{O}(\delta)$ term under Assumption 1 (where $x_k = (x_{1,k}, x_{2,k})$). Hence, (6) can be interpreted as a noisy gradient descent update on a slightly perturbed game. Under Assumptions 1-2 and independent exploration, these dynamics are known to efficiently learn the unique NE.

IV. Bandit Feedback with Deception

The convergence guarantees of exploration-based algorithms, such as (5)-(6), rely on the premise that agents conduct their exploration independently and without interference. However, in competitive, adversarial, or asymmetric environments, these independence and privacy assumptions can break down [15]–[17]. In games, the signals available to the players form an information structure separate from the actions and costs [28]. Models of espionage or information leakage study settings where one player receives noisy or partial information about an opponent's selected strategy or

realized randomization [29]–[31]. These works motivate the need to examine analogous asymmetries in bandit learning, where the leaked randomization is now the mechanism by which each player learns.

Specifically, we consider a setting where player 2 intercepts or infers privileged information regarding player 1's search policy. If player 2 integrates this information into only their own actions, they can manipulate the gradient estimate that player 1 uses to learn. Here, we call player 2 the *deceptive* player, while player 1 is the *oblivious* player. This section formalizes the structure of the leaked information, the DBL dynamics, and their long-term effects on learning.

A. Information Asymmetry

Suppose that at each iteration k, each player selects a search direction $\mu_{i,k} \in \mathbb{S}^n$ independently and uniformly at random, but player 2 also receives a signal $\Theta_{1,k} \in \mathbb{R}^n$ possibly containing privileged information regarding $\mu_{1,k}$. We use $\mathcal{F}_k$ to denote the pre-round history up to stage k, and assume that $\Theta_{1,k}$ satisfies the following structural condition:

Assumption 3. *For $k \in \mathbb{N} \cup \{0\}$, the signal $\Theta_{1,k} \in \mathbb{R}^n$ takes the form*

$$\Theta_{1,k} = \Lambda_k \mu_{1,k} + w_k, \tag{8}$$

where, conditional on $\mathcal{F}_k$, the following hold almost surely:

1) *The directions $\mu_{1,k}$ and $\mu_{2,k}$ are independent and uniformly distributed on $\mathbb{S}^n$. Moreover, the pair (Λ_k, w_k) is conditionally independent of $(\mu_{1,k}, \mu_{2,k})$.*
2) *There exists a matrix $\Gamma \in \mathbb{R}^{n\times n}$, independent of k, such that $\mathbb{E}_k[\Lambda_k] = \Gamma$ a.s, where $\mathbb{E}_k[\,\cdot\,] := \mathbb{E}[\,\cdot\,|\mathcal{F}_k]$.*
3) *There exists constants $\bar{\Lambda}, \bar{w} > 0$ such that $\|\Lambda_k\| \le \bar{\Lambda}$ and $\|w_k\| \le \bar{w}$ a.s.*

All random variables are defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\ge 0}, \mathbb{P})$, with pre-round history given by the σ-algebra $\mathcal{F}_k = \sigma(x_0, \{\mu_{1,\ell}, \mu_{2,\ell}, \Lambda_\ell, w_\ell\}_{\ell<k})$. □

Assumption 3 captures a range of interception scenarios, such as channel leaks, noisy sensing, and intermittent measurements. In a practical context, $\Theta_{1,k}$ represents player 2's imperfect measurement of player 1's exploratory behavior.

The structure of (8) isolates two fundamental sources of imperfection in this observation. The random matrix Λ_k models the transformation, attenuation, or scaling of the intercepted signal. The existence of Γ ensures that the dither-signal cross-covariance is well-behaved. Meanwhile, w_k captures ambient measurement noise or channel interference, which we do not require to be zero mean. More generally, since $\mu_{1,k} = \frac{1}{\delta}(\bar{x}_{1,k} - x_{1,k})$, the signal $\Theta_{1,k}$ may also be a noisy measurement of the played action $\bar{x}_{1,k}$ rather than the actual dither. Notably, the formulation (8) also allows for intermittently available measurements. Indeed, suppose that at stage k the signal is only intercepted with probability $\rho \in (0, 1]$. Taking $\Lambda_k = b_k\widetilde{\Lambda}_k$ and $w_k = b_k\widetilde{w}_k$ with $b_k|\mathcal{F}_k \sim \text{Bernoulli}(\rho)$ (independent of the dithers and $\widetilde{w}_k, \widetilde{\Lambda}_k$) gives $\Gamma = \rho\widetilde{\Gamma}$, where $\widetilde{\Gamma} = \mathbb{E}_k[\widetilde{\Lambda}_k]$. Hence, missed or late interceptions merely attenuate the mean dither-signal cross-covariance by ρ.

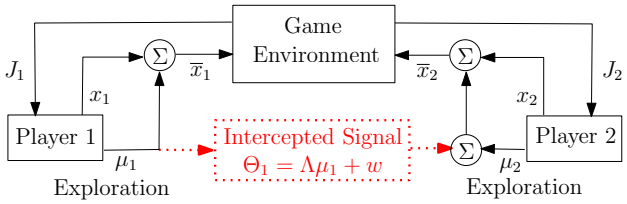


Fig. 1. Exploration leakage and coupling in a two-player game with bandit feedback.

B. Bandit Feedback Algorithm with Deception

With the information asymmetry formally defined, we now construct the modified learning dynamics. We assume player 1 remains oblivious to deception and continues to execute the standard bandit dynamics given below

$$\bar{x}_{1,k} = x_{1,k} + \delta\mu_{1,k} \tag{9a}$$

$$x_{1,k+1} = \text{Proj}_{\mathcal{X}_1^\delta}\left(x_{1,k} - \eta_k \frac{n}{\delta} J_1(\bar{x}_k)\mu_{1,k}\right). \tag{9b}$$

Meanwhile, suppose that player 2 exploits the leaked signal by actively coupling their exploration with player 1's search direction. Before the stage k costs are realized, player 2 injects the signal $\Theta_{1,k}$ into their exploration step with a tunable gain c. The dynamics implemented by player 2 are:

$$\bar{x}_{2,k} = x_{2,k} + \delta\mu_{2,k} + c\delta\Theta_{1,k} \tag{10a}$$

$$x_{2,k+1} = \text{Proj}_{\mathcal{X}_2^{\delta_c}}\left(x_{2,k} - \eta_k \frac{n}{\delta} J_2(\bar{x}_k)\mu_{2,k}\right), \tag{10b}$$

where $\delta_c = \delta(1 + |c|(\bar{\Lambda} + \bar{w}))$, and we assume that δ, c are chosen such that $\delta_c < r_2$. The formulation of δ_c guarantees that even in the worst-case realization of Λ_k and w_k, player 2's exploratory action remains feasible. We will call system (9)-(10) the *deceptive bandit learning (DBL) dynamics*.

The subtlety behind the proposed deception mechanism can be observed through the expected gradient signal, given by $\widehat{g}_{i,k} = \frac{n}{\delta} J_i(\bar{x}_k)\mu_{i,k}$. Computing the smoothed average of $\widehat{g}_{i,k}$ around x_k (details in the Appendix) yields:

$$\mathbb{E}_k[\widehat{g}_{1,k}] = \nabla_1 J_1(x_k) + c\Gamma^\top \nabla_2 J_1(x_k) + \beta_{1,k}^\delta \tag{11a}$$

$$\mathbb{E}_k[\widehat{g}_{2,k}] = \nabla_2 J_2(x_k) + \beta_{2,k}^\delta, \tag{11b}$$

where $\beta_{1,k}^\delta$ and $\beta_{2,k}^\delta$ are $\mathcal{F}_k$-measurable $\mathcal{O}(\delta)$ smoothing biases, with bounds provided in Lemma 2 of the Appendix; such biases are present even in the nominal ($c = 0$) algorithm. The distinctive effect of deception is the induced externality $c\Gamma^\top\nabla_2 J_1(x_k)$, which does not vanish with δ. The system (11) reveals that while player 2 still learns their gradient up to the $\mathcal{O}(\delta)$ bias, player 1 now learns a flawed estimate affected by this externality.

Remark 1: The effect of deception is governed by the conditional cross-covariance between the oblivious player's dither and the intercepted signal:

$$C_{\Theta,k} := n\mathbb{E}_k[\mu_{1,k}\Theta_{1,k}^\top].$$

This correlation introduces the externality $cC_{\Theta,k}\nabla_2 J_1(x_k)$ into the conditional mean of the oblivious player's gradient estimator. Under Assumption 3, $C_{\Theta,k} = \Gamma^\top$ almost surely.

Hence, although the conditional distribution of the signal may adapt to the pre-round history, its fixed cross-covariance produces the same deceptive bias at every iteration. Conversely, if $C_{\Theta,k} = 0$, the intercepted signal cannot generate this equilibrium-shifting bias. This observation also reveals possible defense strategies that can be implemented by the victim, which include desynchronizing the game, applying a randomized zero-mean mask to the dither, etc. An in-depth study of defense strategies is left for future work. □

Remark 2: If players use different exploration radii δ_1, δ_2 with $\delta_i = a_i\delta$, where $a_i > 0$, and (10a) uses $c\delta_2\Theta_{1,k}$, then the externality in (11a) becomes $c(a_2/a_1)\Gamma^\top\nabla_2 J_1(x_k)$, i.e., the effective gain is $c_{\text{eff}} = c(a_2/a_1)$. The computation in (11) also covers the N-player case where player d deceives a player v; dithers are independent, so the injected externality becomes $c\Gamma^\top\nabla_d J_v(x_k)$ and only affects player v's gradient estimator (i.e., the v block of the pseudogradient). □

Motivated by (11), we propose the notion of a *deceptive pseudogradient*, which is given by

$$\mathcal{G}_c^{\text{dec}}(x) = \begin{bmatrix} \nabla_1 J_1(x) + c\Gamma^\top\nabla_2 J_1(x) \\ \nabla_2 J_2(x) \end{bmatrix}. \tag{12}$$

For the modified learning dynamics to converge, $\mathcal{G}_c^{\text{dec}}$ must remain well-behaved. Specifically, the injection of the deceptive externality $c\Gamma^\top\nabla_2 J_1(x)$ must not destroy the monotonicity of the nominal game. By leveraging the Lipschitz continuity of $\nabla_2 J_1(\cdot)$ obtained from Assumption 1, it is straightforward to verify that $\mathcal{G}_c^{\text{dec}}$ remains strongly monotone as long as the following condition is satisfied:

$$\gamma_c := \gamma - |c|L_{12}\|\Gamma\| > 0, \tag{13}$$

where L_{12} is the Lipschitz constant of $\nabla_2 J_1(\cdot)$. If (13) holds, there exists a unique Nash equilibrium of $(\mathcal{G}_c^{\text{dec}}, \mathcal{X})$, which we call the *deceptive Nash equilibrium (DNE)*. This is formalized in the following definition:

Definition 1. *Suppose that $\mathcal{G}_c^{dec}$, given in* (12)*, is strongly monotone. The deceptive Nash equilibrium, denoted x_c^*, is the unique point in $\mathcal{X}$ that satisfies the following*

$$\left\langle \mathcal{G}_c^{dec}(x_c^*),\, x - x_c^* \right\rangle \geq 0, \tag{14}$$

for all $x \in \mathcal{X}$. □

Equations (12), (14) identify a new potential steady state induced by the deception mechanism. This raises two natural questions: do the DBL dynamics approach this DNE, and how does the DNE compare with the nominal NE? We answer these questions next.

V. Main Results

We establish convergence guarantees for the DBL dynamics, then characterize when deception shifts the steady state and benefits the deceiver.

A. Convergence Analysis

Because the coupled gradient estimators now track $\mathcal{G}_c^{\text{dec}}$, the first question is how closely the DBL iterates approach the DNE and at what rate. We provide both guarantees in the following theorem:

Theorem 1. *Consider the DBL dynamics* (9)-(10) *under Assumptions 1-3, where* (13) *holds. Then, there exists $R > 0$ such that, for any $x_0 \in \mathcal{X}_1^\delta \times \mathcal{X}_2^{\delta_c}$ and δ satisfying*

$$0 < \delta < \min\left\{ r_1, \frac{r_2}{1 + |c|(\bar{\Lambda} + \bar{w})} \right\} := r_c, \tag{15}$$

the following bound holds

$$\limsup_{k\to\infty} \mathbb{E}[\|x_k - x_c^*\|^2] \leq R\delta^2. \tag{16}$$

If the step sizes also satisfy $\eta_0 \geq 0$ and $\eta_k = \frac{q}{\gamma_c k}$ for $k \geq 1$, where $q > 1$ and $\gamma_c = \gamma - L_{12}|c|\,\|\Gamma\|$, then there also exists $M > 0$ such that the following holds

$$\mathbb{E}[\|x_k - x_c^*\|^2] \leq \frac{M}{\delta^2 k} + R\delta^2, \tag{17}$$

for all $k \in \mathbb{N}$ and $\delta \in (0, r_c)$.

Thus, while player 1 remains oblivious and continues to execute the nominal bandit update, the joint dynamics converge to a neighborhood of the shifted DNE. From (17), we see that the DBL dynamics inherit the same $\mathcal{O}(d^2/\varepsilon^2)$ complexity guarantees from [18], where $d = 2n = \dim \mathcal{X}$:

Corollary 1. *Under the conditions of Theorem 1 for which* (17) *holds, for any $\varepsilon > 0$ sufficiently small, choosing $\delta^2 = \frac{\varepsilon}{(2R)}$ and $k \geq \frac{4MR}{\varepsilon^2}$ yields $\mathbb{E}[\|x_k - x_c^*\|^2] \leq \varepsilon$, where M, R are given in Section VI-A. Consequently, the DBL dynamics attain the same $\mathcal{O}(d^2/\varepsilon^2)$ complexity as the deception-free bandit algorithm from [18].*

Remark 3: Corollary 1 demonstrates that the dynamics remain efficient even under adversarial manipulation. By treating the smoothing bias in $\mathbb{E}_k[\widehat{G}_k]$ as a bounded additive perturbation (see the proof of Theorem 1 in Section VI-A), we establish these optimal rates while relaxing the assumption of a Lipschitz Jacobian for the pseudogradient, which [18] needed for proving $\mathcal{O}(d^2/\varepsilon^2)$ complexity. The same bounded-perturbation argument extends to N-player games. To our knowledge, our derived relaxation is also novel in the deception-free ($c = 0$) setting. □

Remark 4: Deception is not free. If c satisfies (13), then the strong monotonicity of $\mathcal{G}_c^{\text{dec}}$, with (4) and (14), implies:

$$\|x_c^* - x^*\| \leq \frac{|c|}{\gamma_c}\left\|\Gamma^\top\nabla_2 J_1(x^*)\right\|. \tag{18}$$

So while larger $|c|$ permits larger equilibrium displacement and potential long-term expected cost improvement, it also shrinks γ_c, contracts player 2's feasible set $\mathcal{X}_2^{\delta_c}$, and inflates the smoothing bias bounds in (29); this effectively degrades the constants M and R in Theorem 1. Hence, selecting c itself presents a design trade-off for the deceiver. □

B. Characterization of the DNE and Immunity Conditions

We established convergence to a neighborhood of the DNE, which is characterized by the variational inequality (14). It is evident from (14) that the steady state conditions

are altered under deception, but a fundamental question remains: when is the DNE actually different from the nominal NE? Equation (18) bounds the magnitude of this shift, but we now ask when it is nonzero. Furthermore, in what situations does deception actually improve the cost of the deceiver?

To isolate the effect of the deception from the complexities of active constraints, our vulnerability characterization focuses on interior nominal NE, which is a practical operating condition in many real-world settings [32]–[34]. Specifically, we find that interior NE are locally steerable via deception whenever the externality $\Gamma^\top \nabla_2 J_1(\cdot)$ is nonzero at the nominal NE. Conversely, the equilibrium is invariant whenever $\Gamma^\top \nabla_2 J_1(x^*) = 0$, even if x^* is not necessarily an interior NE. Thus, the zero-correlation condition $\Gamma = 0$ from Remark 1 provides a sufficient universal defense against this particular deception mechanism. Meanwhile, nonzero Γ can still yield immunity if $\mathrm{range}(\Gamma) \perp \nabla_2 J_1(x^*)$, i.e., when the leakage carries no information about the externality. We formalize these ideas in the following theorem:

Theorem 2. *Suppose that Assumptions 1 and 3 hold. If $x^* \in$ int $\mathcal{X}$ and $\Gamma^\top \nabla_2 J_1(x^*) \neq 0$, then the mapping $c \mapsto x_c^*$ is injective on a neighborhood of 0. If instead $\Gamma^\top \nabla_2 J_1(x^*) = 0$, then $x_c^* = x^*$ for all c where $\mathcal{G}_c^{dec}$ is strongly monotone.*

Remark 5: The two immunity conditions in Theorem 2 differ in robustness: $\Gamma = 0$ depends only on the leakage channel and is enforceable by the oblivious player via Remark 1, while $\mathrm{range}(\Gamma) \perp \nabla_2 J_1(x^*)$ depends on the nominal NE and is, hence, unreliable as a defense. □

C. Impact of Deception

While Theorem 2 tells us when the deceiver can steer the dynamics to a new steady state, it is common in practical scenarios that the deceiver is interested in improving their own well-being. If the mapping $c \to x_c^*$ is differentiable around 0, we can evaluate the local sensitivity of J_2 to c. Thus, the question of whether deception is profitable naturally reduces to characterizing this derivative and identifying when it does not vanish. These conditions are formalized below:

Theorem 3. *Suppose that Assumptions 1 and 3 hold, where $\mathcal{G}_c^{dec}$ is continuously differentiable and $x^* \in$ int $\mathcal{X}$. If the following holds:*

$$\begin{bmatrix} \nabla_1 J_2(x^*) \\ 0 \end{bmatrix}^\top \nabla \mathcal{G}(x^*)^{-\top} \begin{bmatrix} \Gamma^\top \nabla_2 J_1(x^*) \\ 0 \end{bmatrix} \neq 0, \tag{19}$$

there exists a nonempty open interval $\mathcal{E} \subset \mathbb{R}$ such that, for all $c \in \mathcal{E}$, $\mathcal{G}_c^{dec}$ is strongly monotone and $J_2(x_c^*) < J_2(x^*)$.

Theorem 3 provides some important geometric intuition behind the mechanics of an adversarially favorable shift. Naturally, for (19) to hold, the term $\Gamma^\top \nabla_2 J_1(x^*)$ cannot be zero. This parallels Theorem 2, which established that this nonzero externality was needed for injectivity of the mapping $c \mapsto x_c^*$. Without this condition, the DNE remains at x^*, which implies that the steady state cost function values will also be unchanged. Surprisingly, condition (19) also requires that the other cross derivative $\nabla_1 J_2(\cdot)$ is nonzero at the NE. Intuitively, this makes sense: if the deceiver's cost is completely insensitive to the victim's actions at the nominal equilibrium, shifting the victim's steady-state behavior yields no immediate benefit. Theorem 1 then tells us that the cost improvements in Theorem 3 can be realized in the long-term expected DBL trajectories whenever δ is sufficiently small.

VI. Proofs of Main Results

A. Proof of Theorem 1

Let $\mathcal{X}_c^\delta := \mathcal{X}_1^\delta \times \mathcal{X}_2^{\delta_c}$, and let $x_{c,\delta}^*$ denote the unique Nash equilibrium of $(\mathcal{G}_c^{\text{dec}}, \mathcal{X}_c^\delta)$. Define $\widehat{G}_k = \frac{n}{\delta}(J_1(\bar{x}_k)\mu_{1,k}, J_2(\bar{x}_k)\mu_{2,k})$. Letting $e_k = x_k - x_{c,\delta}^*$, for each $k \geq 0$ we have:

$$\mathbb{E}[\|x_k - x_c^*\|^2] \leq 2\mathbb{E}\left[\|e_k\|^2\right] + 2\left\|x_{c,\delta}^* - x_c^*\right\|^2. \tag{20}$$

By construction, x_k and e_k are $\mathcal{F}_k$-measurable for every k. To upper bound (20), let us consider the first term, since the second term can be upper bounded with Lemma 1 in the Appendix. Since $x_{c,\delta}^* \in \mathcal{X}_c^\delta$, we can leverage the non-expansiveness of the projection operator to obtain:

$$\begin{aligned}\mathbb{E}_k\left[\|e_{k+1}\|^2\right] &\leq \mathbb{E}_k\left[\left\|x_k - \eta_k \widehat{G}_k - x_{c,\delta}^*\right\|^2\right] \\ &\leq \mathbb{E}_k\left[\|e_k\|^2\right] + \eta_k^2 \mathbb{E}_k\left[\left\|\widehat{G}_k\right\|^2\right] - 2\eta_k\left\langle \mathbb{E}_k\left[\widehat{G}_k\right], e_k\right\rangle.\end{aligned}$$

Let $\Delta_k^\delta = (\beta_{1,k}^\delta, \beta_{2,k}^\delta)$, where $\beta_{1,k}^\delta$ and $\beta_{2,k}^\delta$ are from (11) and satisfy the bounds given in Lemma 2 from the Appendix. Let $\mathcal{J} = \sup_{i\in\{1,2\}} \sup_{x\in\mathcal{X}} |J_i(x)|$, which is finite. Since $\mathbb{E}_k[\widehat{G}_k] = \mathcal{G}_c^{\text{dec}}(x_k) + \Delta_k^\delta$, we obtain:

$$\begin{aligned}\mathbb{E}_k\left[\|e_{k+1}\|^2\right] \leq \mathbb{E}_k\left[\|e_k\|^2\right] &- 2\eta_k \mathbb{E}_k\left[\left\langle \mathcal{G}_c^{\text{dec}}(x_k), e_k\right\rangle\right] \\ &+ \frac{2n^2\eta_k^2\mathcal{J}^2}{\delta^2} - 2\eta_k \mathbb{E}_k\left[\left\langle \Delta_k^\delta, e_k\right\rangle\right] \text{ a.s.}\end{aligned}$$

By the definition of e_k and $x_{c,\delta}^*$, we know that the following holds for all $x_k \in \mathcal{X}_c^\delta$:

$$\left\langle \mathcal{G}_c^{\text{dec}}(x_k), e_k\right\rangle \geq \gamma_c \|e_k\|^2. \tag{21}$$

We can also apply Young's inequality to obtain

$$\left|\left\langle \Delta_k^\delta, e_k\right\rangle\right| \leq \frac{\gamma_c}{2}\|e_k\|^2 + \frac{1}{2\gamma_c}\left\|\Delta_k^\delta\right\|^2. \tag{22}$$

By Lemma 2 in the Appendix, there exists $\alpha > 0$ such that $\left\|\Delta_k^\delta\right\|^2 \leq \alpha\delta^2$ almost surely for each k. We can then combine (21) with (22) and take the total expectation over $\mathcal{F}_k$ to obtain

$$\mathbb{E}\left[\|e_{k+1}\|^2\right] \leq (1 - \eta_k\gamma_c)\mathbb{E}\left[\|e_k\|^2\right] + \frac{2n^2\eta_k^2\mathcal{J}^2}{\delta^2} + \frac{\alpha\eta_k\delta^2}{\gamma_c}. \tag{23}$$

A standard stochastic approximation comparison [35] gives:

$$\limsup_{k\to\infty} \mathbb{E}\left[\left\|x_k - x_{c,\delta}^*\right\|^2\right] \leq \frac{\alpha\delta^2}{\gamma_c^2}. \tag{24}$$

We combine (20), (24), and (28) from Lemma 1 in the Appendix to prove (16), with $R = \frac{2\alpha}{\gamma_c^2} + \frac{2}{r_c^2}(1 + \frac{L_c}{\gamma_c})^2\|x_c^* - p\|^2$. To establish (17), we apply Lemma 3 in the Appendix to (23)

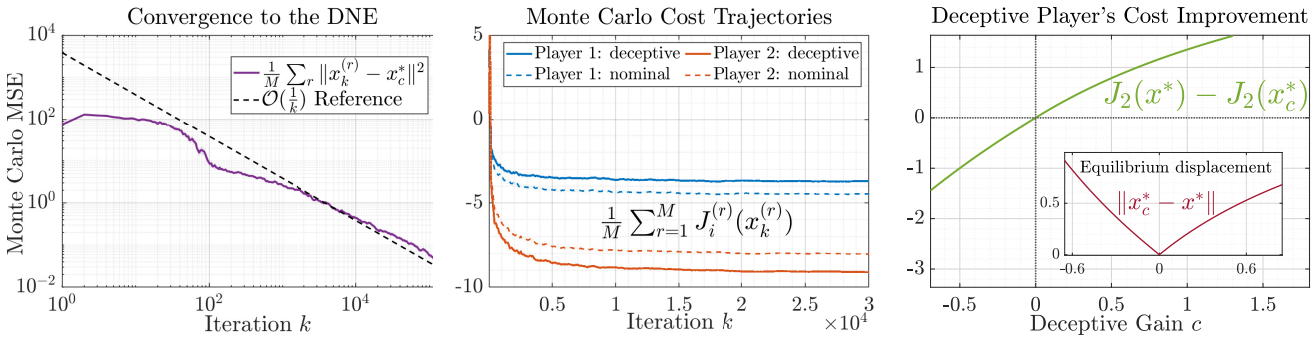


Fig. 2. Monte Carlo study over $M = 300$ realizations. Left: convergence to a neighborhood of the DNE. Middle: nominal and deceptive cost trajectories. Right: the deceptive player's steady-state cost change as a function of c, and the equilibrium displacement.

and obtain $C > 0$ such that the following estimate holds for all $k \in \mathbb{N}$ and $\delta \in (0, r_c)$:

$$\mathbb{E}\left[\left\|x_k - x^*_{c,\delta}\right\|^2\right] \leq \frac{2n^2 C \mathcal{J}^2}{\gamma_c^2 \delta^2 k} + \frac{\alpha \delta^2}{\gamma_c^2}. \tag{25}$$

We substitute (25) into (20) and apply Lemma 1 from the Appendix to (20) and obtain (17), with $C = \max\left\{\frac{q^2}{q-1}, \frac{\gamma_c^2 r_c^2 \lceil q \rceil \max_{x,y\in\mathcal{X}}\|x-y\|^2}{2n^2\mathcal{J}^2}\right\}$, $M = \frac{4Cn^2\mathcal{J}^2}{\gamma_c^2}$. ■

B. Proof of Theorem 2

We start by proving the first part. Since $x^* \in \operatorname{int} \mathcal{X}$, we know that $\mathcal{G}(x^*) = 0$. Let $\varepsilon = \frac{\gamma}{L_{12}\|\Gamma\|}$, and we define the mapping $g : (-\varepsilon, \varepsilon) \to \mathcal{X}$ given by $g(c) = x_c^*$. Since $x^* \in \operatorname{int} \mathcal{X}$ and (18) holds for $c \in (-\varepsilon, \varepsilon)$, there exists a neighborhood $\mathcal{E} \subset (-\varepsilon, \varepsilon)$ of 0 such that $g(\mathcal{E}) \subset \operatorname{int} \mathcal{X}$, and hence $\mathcal{G}_c^{\text{dec}}(g(c)) = 0$ for all $c \in \mathcal{E}$. If we consider some $c_1, c_2 \in \mathcal{E}$ with $c_1 \neq c_2$, we have

$$\mathcal{G}(g(c_i)) + c_i \begin{bmatrix} \Gamma^\top \nabla_2 J_1(g(c_i)) \\ 0 \end{bmatrix} = 0, \tag{26}$$

for $i \in \{1, 2\}$. If we suppose that $g(c_1) = g(c_2) = \hat{x}$, we consider the equation (26) for $i = 1$, $i = 2$ and subtract to obtain

$$(c_1 - c_2) \begin{bmatrix} \Gamma^\top \nabla_2 J_1(\hat{x}) \\ 0 \end{bmatrix} = 0. \tag{27}$$

Since $c_1 - c_2 \neq 0$, the equation (27) is only valid if $\Gamma^\top \nabla_2 J_1(\hat{x}) = 0$. This would imply that $\mathcal{G}(\hat{x}) = 0$, so by the uniqueness of the NE, we conclude that $\hat{x} = x^*$. However, we also assumed that $\Gamma^\top \nabla_2 J_1(x^*) \neq 0$, so we arrive at a contradiction. Hence, g is injective on $\mathcal{E}$, which proves the first part. For the second part, $\Gamma^\top \nabla_2 J_1(x^*) = 0$ implies $\mathcal{G}_c^{\text{dec}}(x^*) = \mathcal{G}(x^*)$. Since x^* solves (4), it also solves (14). Strong monotonicity of $\mathcal{G}_c^{\text{dec}}$ then yields $x_c^* = x^*$. ■

C. Proof of Theorem 3

Since $\mathcal{G}$ is strongly monotone, the implicit function theorem yields $\frac{\partial x_c^*}{\partial c}\Big|_{c=0} = -\nabla \mathcal{G}(x^*)^{-\top} \begin{bmatrix} \Gamma^\top \nabla_2 J_1(x^*) \\ 0 \end{bmatrix}$. Since $x^* \in \operatorname{int} \mathcal{X}$, also have $\nabla_2 J_2(x^*) = 0$. A straightforward application of chain rule reveals that the left hand side of (19) precisely equals $-\frac{\partial J_2(x_c^*)}{\partial c}|_{c=0}$. If $\frac{\partial J_2(x_c^*)}{\partial c}|_{c=0} > 0$, we can find $\widetilde{\varepsilon} \in (0, \varepsilon)$ such that the result holds for $\mathcal{E} = (-\widetilde{\varepsilon}, 0)$, where ε is obtained from the proof of Theorem 2. Otherwise, we will have $\mathcal{E} = (0, \widetilde{\varepsilon})$. ■

VII. Numerical Example

To illustrate our results, we apply them to a stylized two-player resource allocation problem. Consider two players allocating demand across two shared resources. Player i has action $x_i \in \mathbb{R}^2$, where the j-th component of their action specifies the amount of demand assigned to resource j. The cost function of player i is expressed in the form

$$J_i(x) = \|x_i\|^2 + x_i^\top x_{3-i} + q_i^\top x_i + \lambda \sum_{j=1}^{2} [x_{ij} - \tau_{ij}]_+^{\frac{5}{2}},$$

where $[\cdot]_+ := \max\{0, \cdot\}$ and the first three terms model private allocation costs, cross-player congestion, and linear allocation value, respectively. The final term is a soft capacity penalty, where τ_{ij} is player i's reference allocation for resource j. Here, we will use $\lambda = 0.12$, $q_1 = (-4, -5)$, $q_2 = (-6, -4)$, $\tau_1 = (\frac{2}{3}, 2)$, $\tau_2 = (\frac{8}{3}, 1)$, and $\mathcal{X}_1 = \mathcal{X}_2 = [0, 5]^2$. This example satisfies Assumption 1 and falls outside the class of second-order smoothness conditions required in [18].

To seek the Nash equilibrium $x^* = ((2/3, 2), (8/3, 1))$, each player uses a learning algorithm based on bandit feedback. Player 1 implements (9), while, at each stage, player 2 obtains noisy measurements of $\mu_{1,k}$ with probability 0.75, which are stored in a signal $\Theta_{1,k} \in \mathbb{R}^2$ that assumes the form given in (8). We have $\Lambda_k = b_k(\mathbb{I}_2 + \Xi_k)$ where $b_k \sim$ Bernoulli(0.75) and, for each $k \in \mathbb{N}$, each entry of the random matrix Ξ_k is sampled independently and uniformly on $[-0.15, 0.15]$. The additive noise satisfies $w_k = 0.1 b_k z_k$, where the components of z_k are independent standard Gaussian random variables truncated to $[-3, 3]$. Player 2 implements the dynamics (10) with $c = 0.8$, and each player uses $\delta = 0.25$ with step sizes $\eta_k = 3k^{-1}$. It is verifiable that $\mathcal{G}_c^{\text{dec}}$ is strongly monotone and the DNE is given by $(x_c^*)_1 \approx (0.48, 1.44)$ and $(x_c^*)_2 \approx (2.76, 1.26)$. Since $J_2(x^*) - J_2(x_c^*) \approx 1.15$, player 2 improves their steady state cost through deception. The averaged trajectories of J_i and x taken over 300 realizations are given in Figure 2, which also illustrates how the parameter c affects the steady-state costs and equilibrium shift. Figure 3 shows that countermeasures that suppress the same-round dither-signal correlation, which include an induced dither-signal delay, zero-mean masking, and dither resetting, recover the nominal learning trajectories and converge to a neighborhood of the NE, while the undefended dynamics remain displaced.

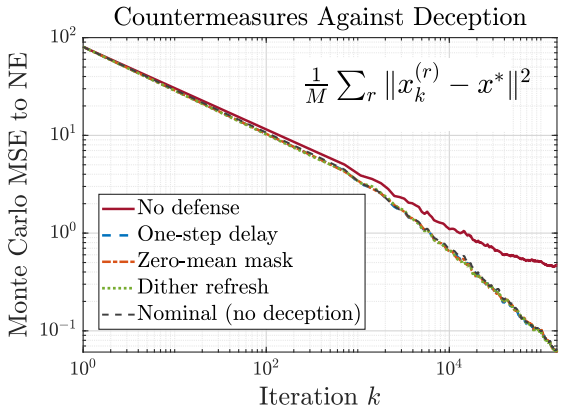


Fig. 3. Convergence of the action profile with deception countermeasures implemented at each stage. One-step delay: player 2 uses $\Theta_{1,k-1}$ during stage k. Zero-mean mask: player 1 privately flips the sign of the dither with probability $\frac{1}{2}$ after leakage. Dither refresh: player 1 resamples the dither after leakage.

VIII. Conclusion

We introduced a novel framework for deceptive bandit learning in strongly monotone games, demonstrating that noisy information correlated with an opponent's exploration can induce an externality that steers the joint action profile toward a deceptive Nash equilibrium. We then proved that the DBL dynamics converge to a $\mathcal{O}(\delta)$ RMS neighborhood of the DNE, establishing optimal rates under weaker regularity conditions than [18] and revealing a novel relaxation that also applies in the deception-free setting. Furthermore, we established geometric conditions under which the deception mechanism is guaranteed to shift the steady state and improve the deceiver's long-term cost/payoff. Our results were illustrated in a two-player resource allocation setting. Future directions include extending the ideas to asynchronous games, studying the effect of the network topology, and investigating active countermeasures against deception.

References

[1] A. Flaxman, A. Kalai, and H. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005, pp. 385–394.

[2] P. Landgren, V. Srivastava, and N. E. Leonard, "Distributed cooperative decision making in multi-agent multi-armed bandits," *Automatica*, vol. 125, p. 109445, 2021.

[3] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE transactions on signal processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[4] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2786–2790.

[5] J. Zhu and J. Liu, "Distributed multiarmed bandits," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 3025–3040, 2023.

[6] A. Sankararaman, A. Ganesh, and S. Shakkottai, "Social learning in multi agent multi armed bandits," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 3, pp. 1–35, 2019.

[7] K. Misra, E. M. Schwartz, and J. Abernethy, "Dynamic online pricing with incomplete information using multiarmed bandit experiments," *Marketing Science*, vol. 38, no. 2, pp. 226–252, 2019.

[8] J. W. Mueller, V. Syrgkanis, and M. Taddy, "Low-rank bandit methods for high-dimensional dynamic pricing," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] J. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, 1992.

[10] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

[11] D. Malik *et al.*, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 2916–2925.

[12] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.

[13] T. Tatarenko and M. Kamgarpour, "Learning generalized Nash equilibria in a class of convex games," *IEEE Transactions on Automatic Control*, vol. 64, no. 4, pp. 1426–1439, 2018.

[14] Y. Pang and G. Hu, "Distributed Nash equilibrium seeking with limited cost function knowledge via a consensus-based gradient-free method," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1832–1839, 2020.

[15] K. G. Vamvoudakis *et al.*, "Deception in game theory and control: A tutorial," in *2025 American Control Conference (ACC)*. IEEE, 2025, pp. 2146–2163.

[16] M. Ye, G. Hu, L. Xie, and S. Xu, "Differentially private distributed Nash equilibrium seeking for aggregative games," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2451–2458, 2021.

[17] E. Boursier and V. Perchet, "Selfish robustness and equilibria in multi-player bandits," in *Conference on Learning Theory*. PMLR, 2020, pp. 530–581.

[18] D. Drusvyatskiy, M. Fazel, and L. J. Ratliff, "Improved rates for derivative free gradient play in strongly monotone games," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 3403–3408.

[19] F. Yao, Y. Cheng, E. Wei, and H. Xu, "Single-agent poisoning attacks suffice to ruin multi-agent learning," in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Y. Wu, J. McMahan, X. Zhu, and Q. Xie, "Reward poisoning attacks on offline multi-agent reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 10 426–10 434.

[21] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," *arXiv preprint arXiv:1905.10615*, 2019.

[22] M. Tang, U. Javed, X. Chen, M. Krstić, and J. I. Poveda, "Deception in Nash equilibrium seeking," *IEEE Transactions on Automatic Control*, vol. 70, no. 12, pp. 7984–7999, 2025.

[23] M. Tang, M. Krstic, and J. Poveda, "Stochastic real-time deception in Nash equilibrium seeking for games with quadratic payoffs," in *7th Annual Learning for Dynamics & Control Conference*. PMLR, 2025, pp. 635–646.

[24] Y.-H. Yan, P. Zhao, and Z.-H. Zhou, "Fast rates in time-varying strongly monotone games," in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 138–39 164.

[25] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave n-person games," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] W. Ba, T. Lin, J. Zhang, and Z. Zhou, "Doubly optimal no-regret online learning in strongly monotone games with bandit feedback," *Operations Research*, vol. 73, no. 6, pp. 3219–3244, 2025.

[27] M. Jordan, T. Lin, and Z. Zhou, "Adaptive, doubly optimal no-regret learning in strongly monotone and exp-concave games with gradient feedback," *Operations Research*, vol. 73, no. 3, pp. 1675–1702, 2025.

[28] D. Bergemann and S. Morris, "Bayes correlated equilibrium and the comparison of information structures in games," *Theoretical Economics*, vol. 11, no. 2, pp. 487–522, 2016.

[29] E. Solan and L. Yariv, "Games with espionage," *Games and Economic Behavior*, vol. 47, no. 1, pp. 172–199, 2004.

[30] N. Alon, Y. Emek, M. Feldman, and M. Tennenholtz, "Adversarial leakage in games," *SIAM Journal on Discrete Mathematics*, vol. 27, no. 1, pp. 363–385, 2013.

[31] M. S. Alvim, K. Chatzikokolakis, Y. Kawamoto, and C. Palamidessi, "Information leakage games," in *International Conference on Decision and Game Theory for Security*. Springer, 2017, pp. 437–457.

[32] P. Frihauf, M. Krstic, and T. Basar, "Nash equilibrium seeking in

noncooperative games," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1192–1207, 2012.

[33] B. Gao and L. Pavel, "Bandit learning with regularized second-order mirror descent," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 5731–5738.

[34] J. Perolat *et al.*, "From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization," in *ICML*. PMLR, 2021, pp. 8525–8535.

[35] B. Polyak, *Introduction to Optimization*, ser. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987.

IX. Appendix

Here we present several technical lemmas that are used to prove our main results.

Lemma 1. *Suppose $\mathcal{G}_c^{dec}$ is γ_c-strongly monotone, and let $x^*_{c,\delta}$ denote a Nash equilibrium of $(\mathcal{G}_c^{dec}, \mathcal{X}_1^\delta \times \mathcal{X}_2^{\delta_c})$, where δ satisfies* (15)*. Then, the following holds*

$$\left\|x_c^* - x_{c,\delta}^*\right\| \leq d_c \left(1 + \frac{L_c}{\gamma_c}\right) \|x_c^* - p\|, \tag{28}$$

where $d_c = \max\left(\frac{\delta}{r_1}, \frac{\delta_c}{r_2}\right)$, $L_c := |c|\,\|\Gamma\|\,L_{12} + \sqrt{L_{11}^2 + L_{22}^2}$, *and* $p = (p_1, p_2)$.

Proof. Define $\mathcal{X}_c^\delta := \mathcal{X}_1^\delta \times \mathcal{X}_2^{\delta_c}$, which is convex and compact. Letting $p = (p_1, p_2)$, we define the following matrix and transformation:

$$\mathcal{D} := \begin{bmatrix} \left(1 - \frac{\delta}{r_1}\right)\mathbb{I}_n & 0 \\ 0 & \left(1 - \frac{\delta_c}{r_2}\right)\mathbb{I}_n \end{bmatrix}, \quad T(x) = \mathcal{D}(x - p) + p.$$

Note that $\mathcal{X}_c^\delta = T(\mathcal{X})$. By the strong monotonicity of $\mathcal{G}_c^{\text{dec}}$, it also follows that the mapping $x \mapsto \mathcal{G}_c^{\text{dec}}(x) + \mathcal{N}_{\mathcal{X}_c^\delta}(x)$ is also γ_c-strongly monotone, where $\mathcal{N}_{\mathcal{X}_c^\delta}(x)$ is the normal cone of $\mathcal{X}_c^\delta$ at the point x. We then obtain

$$\gamma_c \left\|T(x_c^*) - x_{c,\delta}^*\right\| \leq \inf_{v \in \mathcal{G}_c^{\text{dec}}(T(x_c^*)) + \mathcal{N}_{\mathcal{X}_c^\delta}(T(x_c^*))} \|v\|.$$

Since x_c^* is a DNE, we have that $0 \in \mathcal{G}_c^{\text{dec}}(x_c^*) + \mathcal{N}_{\mathcal{X}}(x_c^*)$. Using the fact that $\mathcal{N}_{\mathcal{X}_c^\delta}(T(x_c^*)) = \mathcal{N}_{\mathcal{X}}(x_c^*)$, we obtain

$$\begin{aligned} \inf_{v \in \mathcal{G}_c^{\text{dec}}(T(x_c^*)) + \mathcal{N}_{\mathcal{X}_c^\delta}(T(x_c^*))} \|v\| &= \inf_{v \in \mathcal{G}_c^{\text{dec}}(T(x_c^*)) + \mathcal{N}_{\mathcal{X}}(x_c^*)} \|v\| \\ &\leq \left\|\mathcal{G}_c^{\text{dec}}(T(x_c^*)) - \mathcal{G}_c^{\text{dec}}(x_c^*)\right\| \\ &\leq L_c \left\|(\mathcal{D} - \mathbb{I}_{2n})(x_c^* - p)\right\| \\ &\leq L_c d_c \|x_c^* - p\|, \end{aligned}$$

where L_c is a Lipschitz constant of $\mathcal{G}_c^{\text{dec}}$. By combining the above computations, we obtain

$$\begin{aligned} \left\|x_c^* - x_{c,\delta}^*\right\| &\leq \|x_c^* - T(x_c^*)\| + \left\|T(x_c^*) - x_{c,\delta}^*\right\| \\ &\leq d_c \left(1 + \frac{L_c}{\gamma_c}\right) \|x_c^* - p\|, \end{aligned}$$

as desired. □

Lemma 2. *In* (11)*, for each $k \in \mathbb{N}$, the following estimates hold almost surely:*

$$\left\|\beta_{1,k}^\delta\right\| \leq \delta(L_{11} + |c|\bar{\Lambda}L_{12})(2 + |c|(\bar{\Lambda} + \bar{w})) \tag{29a}$$

$$\left\|\beta_{2,k}^\delta\right\| \leq \delta L_{22}(2 + |c|(\bar{\Lambda} + \bar{w})). \tag{29b}$$

Proof. Given $k \in \mathbb{N}$, we recall that

$$\bar{x}_k = \begin{bmatrix} x_{1,k} + \delta\mu_{1,k} \\ x_{2,k} + \delta\mu_{2,k} + c\delta(\Lambda_k\mu_{1,k} + w_k) \end{bmatrix}. \tag{30}$$

Since k will remain fixed, we drop it from the subscripts for notational convenience. By Stokes' theorem:

$$\mathbb{E}_k\left[\frac{n}{\delta}J_1(\bar{x})\mu_1\right] = \mathbb{E}_k^{\mu_1 \sim B, \mu_2 \sim S}\left[\nabla_1 J_1(\bar{x}) + c\Lambda^\top \nabla_2 J_1(\bar{x})\right]$$

$$\mathbb{E}_k\left[\frac{n}{\delta}J_2(\bar{x})\mu_2\right] = \mathbb{E}_k^{\mu_1 \sim S, \mu_2 \sim B}\left[\nabla_2 J_2(\bar{x})\right],$$

where S and B are uniform distributions over $\mathbb{S}^n$ and $\mathbb{B}^n$, respectively. Since $\mathbb{E}_k[\Lambda] = \Gamma$, we can continue from (11) to obtain

$$\begin{aligned} \left\|\beta_1^\delta\right\| &\leq \left\|\mathbb{E}_k^{\mu_1 \sim B, \mu_2 \sim S}\left[\nabla_1 J_1(\bar{x}) + c\Lambda^\top \nabla_2 J_1(\bar{x})\right]\right. \\ &\quad \left. - \nabla_1 J_1(x) - c\Gamma^\top \nabla_2 J_1(x)\right\| \\ &\leq \mathbb{E}_k^{\mu_1 \sim B, \mu_2 \sim S}\left[\,\|\nabla_1 J_1(\bar{x}) - \nabla_1 J_1(x)\|\right. \\ &\quad \left. + \left\|c\Lambda^\top(\nabla_2 J_1(\bar{x}) - \nabla_2 J_1(x))\right\|\,\right] \\ &\leq \mathbb{E}_k^{\mu_1 \sim B, \mu_2 \sim S}\left[(L_{11} + |c|\,\|\Lambda\|\,L_{12})\,\|\bar{x} - x\|\right], \end{aligned}$$

and

$$\begin{aligned} \left\|\beta_2^\delta\right\| &\leq \mathbb{E}_k^{\mu_1 \sim S, \mu_2 \sim B}\left[\|\nabla_2 J_2(\bar{x}) - \nabla_2 J_2(x)\|\right] \\ &\leq \mathbb{E}_k^{\mu_1 \sim S, \mu_2 \sim B}\left[L_{22}\,\|\bar{x} - x\|\right]. \end{aligned}$$

The result is obtained by substituting (30) into the above expressions and using the facts that $\|\mu_i\| \leq 1$ and $\|\Lambda\| \leq \bar{\Lambda}$ almost surely. □

Lemma 3. *Consider a nonnegative sequence $\{V_k\}_{k=1}^\infty$ satisfying $V_{k+1} \leq (1 - a_k)V_k + Aa_k^2 + Ba_k$, where $a_k = \frac{q}{k}$ for some $q > 1$, and $A, B > 0$. Then, there exists $C > 0$ such that $V_k \leq \frac{CA}{k} + B$ for all $k \in \mathbb{N}$.*

Proof. The desired bound clearly holds for $k \leq \lceil q \rceil$, since we can always choose $C \geq \frac{1}{A}\max_{k \leq \lceil q \rceil} kV_k$, so suppose it holds for some $k \geq \lceil q \rceil$. By the inductive hypothesis, we have $V_{k+1} \leq \frac{CA(k-q) + Aq^2}{k^2} + B$. It suffices to show $\frac{CA(k-q)+Aq^2}{k^2} \leq \frac{CA}{k+1}$, which can be equivalently rearranged into $k\left(CA(1-q) + Aq^2\right) - CAq + Aq^2 \leq 0$. Direct computation shows that this holds for all $k \in \mathbb{N}$ when $C \geq \frac{q^2}{q-1}$. We set $C \geq \max\left\{\frac{q^2}{q-1}, \frac{1}{A}\max_{k \leq \lceil q \rceil} kV_k\right\}$ to conclude. □