Decentralized Learning of Belief-Dependent Generalized Nash Equilibria under Partial Observability

Frédy Pokou
Inria, University of Lille, CNRS, Centrale Lille
Villeneuve-d'Ascq, France
fredy-vale-manuel.pokou@inria.fr

Hélène Le Cadre
Inria, University of Lille, CNRS, Centrale Lille
Villeneuve-d'Ascq, France
helene.le-cadre@inria.fr

ABSTRACT

We study distributed decision making under partial observability in constrained multi-agent systems. Agents operate with local sensing, limited communication, and noisy exogenous information while jointly satisfying coupled proximity constraints and learning an unknown latent parameter governing payoffs. We model the interaction as a belief-dependent generalized Nash equilibrium (GNE), defined as the equilibrium of an augmented belief-state game induced by agents' posterior beliefs. We develop a fully decentralized discrete-time learning scheme combining quadratic-penalty constraint enforcement with distributed Thompson sampling, enabling agents to simultaneously learn the payoff parameter and coordinate on constraint-feasible equilibria. Under mild regularity, constraint qualification, and posterior concentration assumptions, we prove that the induced stochastic dynamics converges almost surely to a belief-dependent GNE of the true constrained game. Finite-time performance is characterized through information-theoretic Bayesian regret bounds. Numerical experiments validate the theory and demonstrate that information quality, rather than information quantity, is the dominant driver of coordination.

KEYWORDS

Generalized Nash Equilibrium, Incomplete Information Games, Decentralized Multi-Agent Learning

ACM Reference Format:
Frédy Pokou and Hélène Le Cadre. 2026. Decentralized Learning of Belief-Dependent Generalized Nash Equilibria under Partial Observability. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages.

1 INTRODUCTION

Classical Lanchester-type combat models [12] and their asymmetric extensions provide early analytical foundations for adversarial multi-agent interaction. Subsequent formulations incorporated mobility, dispersion, and heterogeneous forces, leading to guerrilla warfare models [2] and fire-allocation games [6]. While these models yield tractable equilibrium characterizations, they typically assume centralized or perfectly aggregated information and rely on static decision rules, limiting their ability to capture coordination under decentralized information.

Operations research has emphasized the role of intelligence and situational awareness in adversarial decision-making [8, 9]. However, information is usually treated as an exogenous parameter rather than an object learned and coordinated upon by agents operating under uncertainty. From a multi-agent perspective, this abstraction overlooks how decentralized information acquisition and sharing shape equilibrium behavior and collective performance.

More generally, agents in networked systems must simultaneously learn unknown parameters, anticipate others' actions, and satisfy shared feasibility constraints using only local information. Understanding how such information structures affect equilibrium selection, coordination efficiency, and social welfare remains a central challenge at the intersection of multi-agent learning and game theory, motivating frameworks that jointly couple belief formation, strategic decision making, and constrained coordination.

In parallel, the learning and control literature has developed posterior sampling and regret-minimization tools for partially observable environments. Thompson sampling achieves optimal regret guarantees in single-agent and competitive reinforcement learning [1, 14–16, 19, 20]. Extensions to multi-agent settings include communication-efficient Thompson sampling [7], posterior-sampling exploration [22], and no-regret learning in general-sum Markov games [4]. Complementary work studies decentralized equilibrium-seeking dynamics [3, 5] and MARL environments [18, 21, 22]. Nevertheless, these approaches largely abstract away spatially localized sensing, heterogeneous exogenous information sources, and coordination inefficiencies arising purely from decentralized information structures.

Despite recent progress, a unified framework capturing decentralized information acquisition, equilibrium coordination under coupled constraints, and Bayesian learning of unknown payoff parameters remains largely unexplored. Existing game-theoretic learning models often assume known payoffs or centralized information, while decentralized learning approaches focus on unconstrained environments or ignore equilibrium consistency. Consequently, the interaction between learning, information decentralization, and constraint-coupled strategic behavior remains insufficiently understood.

This paper makes three main contributions.
(i) **Framework.** We introduce a game-theoretic model for decentralized equilibrium learning under endogenous information acquisition, formulating agent interaction as a belief-dependent Generalized Nash Equilibrium (GNE) arising from an augmented belief-state game with coupled feasibility constraints. This framework captures the joint effects of partial observability, heterogeneous information sources, and decentralized coordination.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.

(ii) **Algorithm.** We propose a fully decentralized discrete-time learning algorithm that integrates quadratic-penalty equilibrium-seeking dynamics with distributed Thompson sampling, enabling agents to simultaneously learn unknown payoff parameters and coordinate on constraint-feasible equilibria using only local sensing and limited communication.
(iii) **Guarantees and metrics.** We establish almost sure convergence of the induced stochastic dynamics to a Generalized Nash Equilibrium of the true constrained game and derive sublinear Bayesian regret guarantees. We further introduce the Price of Miscoordination (PoM), a welfare-based metric that isolates coordination losses induced purely by decentralized information and complements regret-based performance analysis.

2 GAME ENVIRONMENT, INFORMATION, AND PAYOFFS

Let $\mathcal{N} := \{1, \dots, N\}$ denote the set of agents. Each agent $n \in \mathcal{N}$ selects a continuous action $x_n \in \mathcal{X}_n \subset \mathbb{R}^d$, where $\mathcal{X}_n$ is nonempty, compact and convex. The joint action profile is $\mathbf{x} := (x_1, \dots, x_N) \in \mathcal{X} := \prod_{n\in\mathcal{N}} \mathcal{X}_n$.

Agents interact locally over an undirected connected graph

$$G := (\mathcal{N}, E), \quad \mathcal{V}_n := \{n' \in \mathcal{N} : \{n, n'\} \in E\}.$$

Each agent observes only neighbors' actions $x_{n'}$ for $n' \in \mathcal{V}_n$, which induces local strategic coupling and decentralized information flow.

The environment depends on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^K$ drawn once and fixed over time. Agents do not observe θ directly and instead maintain subjective beliefs over Θ.

Let $\pi_n \in \Delta(\Theta)$ denote agent n's belief distribution over θ, where $\Delta(\Theta)$ denotes the space of probability measures over Θ.

Agent n evaluates the joint action profile $\mathbf{x}$ through the payoff

$$f_n(x_n, \mathbf{x}_{-n}; \theta, \mu) = U_n(x_n, \mathbf{x}_{-n}; \theta, \mu) - C_n(x_n), \tag{1}$$

where U_n denotes a continuously differentiable interaction utility capturing strategic coupling with neighboring agents and dependence on the latent environment parameter θ, while C_n is a continuously differentiable individual cost function. We assume that C_n is convex and that U_n is jointly continuous in all arguments and continuously differentiable in $(x_n, \mathbf{x}_{-n})$.

The payoff admits an additive decomposition into an interaction component $U_n(\cdot; \theta, \mu)$ and a separable individual cost $C_n(\cdot)$. Under symmetric interaction neighborhoods and standard regularity conditions, the induced game admits an exact potential function given by the aggregate interaction utility minus the sum of individual costs. This form of payoffs is specific to coordination games, e.g. in [10]. This structure is consistent with gradient-based decentralized equilibrium-seeking dynamics and facilitates the analysis of convergence under learning and stochastic perturbations.

Agents may receive signals from non-strategic exogenous information sources. The parameter $\bar{\rho} \in [0, \infty)$ specifies the sensing or communication radius determining which sources are locally observable by agent n, while the trust parameter $\mu \in [0, 1]$ captures the intrinsic informativeness (or reliability) of these signals and determines how they enter the interaction utility U_n. Hence, $\bar{\rho}$ controls information availability and μ controls information quality, independently of the interaction radius $\rho \in [0, \infty)$ governing feasibility constraints.

Throughout the paper, the parameter μ is assumed to be fixed and common knowledge, and therefore remains constant over the time horizon of the game. For notational simplicity, and since no learning or strategic adaptation occurs with respect to μ, we omit its explicit dependence in the payoff functions and write $f_n(x_n, \mathbf{x}_{-n}; \theta)$ and $U_n(x_n, \mathbf{x}_{-n}; \theta)$ whenever no ambiguity arises.

Assumption 1 (Payoff regularity). *For every $\theta \in \Theta$ and $n \in \mathcal{N}$:*
(1) *$f_n(\cdot, \mathbf{x}_{-n}; \theta)$ is concave in x_n,*
(2) *$\nabla_{x_n} f_n(\cdot, \mathbf{x}_{-n}; \theta)$ is Lipschitz continuous,*
(3) *f_n is continuously differentiable in $(\mathbf{x}, \theta)$.*

These conditions ensure well-posed best-responses and enable variational characterizations of equilibria.

Let $0 < \rho < \infty$ be a parameter defining the communication range of the agents. Actions must satisfy pairwise proximity constraints

$$g_{n,n'}(x_n, x_{n'}) := \|x_n - x_{n'}\| - \rho \le 0, \quad \forall n \in \mathcal{N}, \quad \forall n' \in \mathcal{V}_n.$$

The joint feasible set is

$$\mathcal{X}_c := \left\{\mathbf{x} \in \mathcal{X} : g_{n,n'}(x_n, x_{n'}) \le 0,\ \forall \{n, n'\} \in E\right\}.$$

Since $\mathcal{X}_c$ is defined by pairwise edge constraints, the section $\mathcal{X}_n(\mathbf{x}_{-n}) = \{x_n \in \mathcal{X}_n : (x_n, \mathbf{x}_{-n}) \in \mathcal{X}_c\}$ reduces to

$$\mathcal{X}_n(\mathbf{x}_{-n}) = \{x_n \in \mathcal{X}_n :\ g_{n,n'}(x_n, x_{n'}) \le 0,\ \forall n' \in \mathcal{V}_n\},$$

since constraints associated with edges not incident to n do not depend on x_n.

Assumption 2 (Constraint qualification). *At any feasible point where $g_{n,n'}(x_n, x_{n'}) = 0$ and $x_n \neq x_{n'}$, the gradients $\nabla g_{n,n'}$ are nonzero.*

2.1 Belief-Dependent Generalized Nash Game

The environment is characterized by a fixed but unknown parameter $\theta \in \Theta$. The true interaction therefore corresponds to the generalized Nash game (GNG)

$$\mathcal{G}(\theta) := \Big(\mathcal{N},\ (\mathcal{X}_n)_{n\in\mathcal{N}},\ \mathcal{X}_c,\ (f_n(\cdot; \theta))_{n\in\mathcal{N}}\Big). \tag{2}$$

Since agents do not know θ, they instead play a sequence of belief-dependent GNGs

$$\mathcal{G}(\boldsymbol{\pi}^{(t)}) := \Big(\mathcal{N},\ (\mathcal{X}_n)_{n\in\mathcal{N}},\ \mathcal{X}_c,\ (\bar{f}_n(\cdot; \pi_n^{(t)}))_{n\in\mathcal{N}}\Big),$$

where $\bar{f}_n(x_n, \mathbf{x}_{-n}; \pi_n^{(t)})$ is the belief-dependent payoff

$$\bar{f}_n(x_n, \mathbf{x}_{-n}; \pi_n^{(t)}) = \mathbb{E}_{\theta\sim\pi_n^{(t)}}\left[f_n(x_n, \mathbf{x}_{-n}; \theta)\right]. \tag{3}$$

2.1.1 Time-Varying Game Formulation. Time is discrete $t \in \mathbb{N}$. At each time step, agents play a *stage generalized Nash game* induced by the current information state (beliefs and observations) and the active set $\mathcal{N}_n^{(t)}$. We denote the stage payoff by $f_n^{(t)}(x_n, \mathbf{x}_{-n}; \theta)$ when emphasizing time dependence; when no ambiguity arises the superscript (t) is omitted for readability. The learning algorithm defined later therefore induces a stochastic process over the sequence of stage games and the joint state $(\mathbf{x}^{(t)}, \pi_n^{(t)}, \lambda^{(t)})$. Because the feasible set of each agent depends on the actions of its neighbors through (3), the interaction is a generalized Nash game rather than a standard Nash game and $\lambda^{(t)}$ is the dual variable of the coupling constraints.

2.1.2 Solution Concepts.

Definition 1 (Generalized Nash Equilibrium). *A joint action $\mathbf{x}^\star \in \mathcal{X}_c$ is a generalized Nash equilibrium (GNE) if no agent can improve its utility by a unilateral feasible deviation, i.e.,*

$$f_n(x_n^\star, \mathbf{x}_{-n}^\star; \theta) \geq f_n(x_n, \mathbf{x}_{-n}^\star; \theta) \quad \forall x_n \in \mathcal{X}_n(\mathbf{x}_{-n}^\star),\ \forall n \in \mathcal{N}.$$

Among GNEs, we focus on equilibria compatible with a shared set of constraint multipliers leading to the same dual variables λ.

Definition 2 (Variational GNE). *A GNE $\mathbf{x}^\star$ is called a variational GNE (v-GNE) if it satisfies the first-order optimality conditions of a common variational inequality over the collective feasible set $\mathcal{X}_c$. Equivalently, all agents share consistent dual variables associated with the coupling constraints.*

We extend these definitions to the belief-dependent GNG setting.

Definition 3 (Belief-dependent generalized Nash equilibrium). *Fix a belief profile $\boldsymbol{\pi} := (\pi_n)_{n\in\mathcal{N}}$. A joint action $\mathbf{x}^\star \in \mathcal{X}_c$ is a generalized Nash equilibrium (GNE) of the belief-dependent game $\mathcal{G}(\boldsymbol{\pi})$ if for every agent $n \in \mathcal{N}$,*

$$x_n^\star \in \arg\max_{x_n \in \mathcal{X}_n(\mathbf{x}_{-n}^\star)} \bar{f}_n(x_n, \mathbf{x}_{-n}^\star; \pi_n), \tag{4}$$

Definition 4 (Belief-dependent variational generalized Nash equilibrium). *A point $\mathbf{x}^\star \in \mathcal{X}_c$ is a variational generalized Nash equilibrium (v-GNE) of $\mathcal{G}(\boldsymbol{\pi})$ if it solves the variational inequality*

$$\langle F(\mathbf{x}^\star; \boldsymbol{\pi}), \mathbf{x} - \mathbf{x}^\star \rangle \leq 0, \quad \forall \mathbf{x} \in \mathcal{X}_c, \tag{5}$$

where the pseudo-gradient mapping is

$$F(\mathbf{x}; \boldsymbol{\pi}) := \big(\nabla_{x_n} \bar{f}_n(\mathbf{x}; \pi_n)\big)_{n\in\mathcal{N}}.$$

In practice, agents implement a posterior-sampling (Thompson-style [1, 17]) decision rule. At each time step t, agent n samples a parameter realization $\theta_n^{(t)} \sim \pi_n^{(t)}$ and computes a best-response under the sampled model

$$x_n^{(t+1)} \in \arg\max_{x_n \in \mathcal{X}_n(\mathbf{x}_{-n}^{(t)})} f_n(x_n, \mathbf{x}_{-n}^{(t)}; \theta_n^{(t)}). \tag{6}$$

Hence, while equilibrium analysis is naturally formulated in the belief game $\mathcal{G}(\boldsymbol{\pi}^{(t)})$, the implemented dynamics correspond to stochastic best-responses generated by sampled games $\mathcal{G}(\theta_n^{(t)})$, $\forall n$.

2.2 Potential Game

We next investigate whether the game admits a potential structure, which is useful for equilibrium characterization and learning convergence analysis.

Definition 5 (Potential game). *The complete-information game $\mathcal{G}(\theta)$ is a potential game if there exists a continuously differentiable function $\Phi(\mathbf{x}; \theta)$ such that, for all $n \in \mathcal{N}$,*

$$\nabla_{x_n} f_n(x_n, \mathbf{x}_{-n}; \theta) = \nabla_{x_n} \Phi(\mathbf{x}; \theta), \quad \forall \mathbf{x} \in \mathcal{X}_c.$$

In this case, v-GNEs of $\mathcal{G}(\theta)$ coincide with optima of $\Phi(\cdot; \theta)$ over $\mathcal{X}_c$.

Define the belief-averaged objective

$$\bar{\Phi}(\mathbf{x}; \boldsymbol{\pi}) := \sum_{n\in\mathcal{N}} \mathbb{E}_{\theta\sim\pi_n}\big[f_n(x_n, \mathbf{x}_{-n}; \theta)\big] \tag{7}$$

defines a heterogeneous potential, which preserves alignment of unilateral incentives but does not, in general, coincide with the expectation of a common potential function.

Proposition 1 (Potential structure). *Fix $\theta \in \Theta$ and consider $\mathcal{G}(\theta)$ with utilities $f_n(x_n, \mathbf{x}_{-n}; \theta)$ and feasible set $\mathcal{X}_c$. Assume that*

$$f_n(x_n, \mathbf{x}_{-n}; \theta) = u_n(x_n; \theta) + \sum_{n'\in\mathcal{V}_n} \psi_{n,n'}(x_n, x_{n'}; \theta) - C_n(x_n), \tag{8}$$

where for all $\{n, n'\} \in E$, $\psi_{n,n'}(x_n, x_{n'}; \theta) = \psi_{n',n}(x_{n'}, x_n; \theta)$. Then $\mathcal{G}(\theta)$ is an exact potential generalized Nash game with potential

$$\Phi(\mathbf{x}; \theta) = \sum_{n\in\mathcal{N}} f_n(x_n, \mathbf{x}_{-n}; \theta). \tag{9}$$

Proof. Let $\mathbf{x} \in \mathcal{X}_c$ and fix $\theta \in \Theta$. Using the separable structure,

$$f_n(x_n, \mathbf{x}_{-n}; \theta) = u_n(x_n; \theta) + \sum_{n'\in\mathcal{V}_n} \psi_{n,n'}(x_n, x_{n'}; \theta) - C_n(x_n).$$

Define $\Phi(\mathbf{x}; \theta) = \sum_{n\in\mathcal{N}} f_n(x_n, \mathbf{x}_{-n}; \theta)$. Consider any unilateral deviation $x_n \to x_n'$, keeping $\mathbf{x}_{-n}$ fixed. Only the terms involving agent n change in the potential. By symmetry of pairwise interactions, $\psi_{n,n'}(x_n, x_{n'}; \theta) = \psi_{n',n}(x_{n'}, x_n; \theta)$, so each pairwise term appears exactly once in Φ. Hence, $\Phi(x_n', \mathbf{x}_{-n}; \theta) - \Phi(x_n, \mathbf{x}_{-n}; \theta) = f_n(x_n', \mathbf{x}_{-n}; \theta) - f_n(x_n, \mathbf{x}_{-n}; \theta)$. Therefore, Φ is an exact potential for $\mathcal{G}(\theta)$ on $\mathcal{X}_c$, which proves the result. □

Proposition 2 (Belief-dependent potential game with heterogeneous beliefs). *Assume that for every $\theta \in \Theta$, the game $\mathcal{G}(\theta)$ is an exact potential game with potential $\Phi(\mathbf{x}; \theta)$, i.e.,*

$$\nabla_{x_n} f_n(x_n, \mathbf{x}_{-n}; \theta) = \nabla_{x_n} \Phi(\mathbf{x}; \theta), \quad \forall n \in \mathcal{N}.$$

Then, the belief-dependent game $\mathcal{G}(\boldsymbol{\pi})$ admits a generalized (heterogeneous) potential structure in the sense that, for every $n \in \mathcal{N}$,

$$\nabla_{x_n} \bar{f}_n(\mathbf{x}; \pi_n) = \nabla_{x_n} \bar{\Phi}(\mathbf{x}; \boldsymbol{\pi}),$$

where $\bar{f}_n(\mathbf{x}; \pi_n) = \mathbb{E}_{\theta\sim\pi_n}[f_n(\mathbf{x}; \theta)]$.

Remark 1. *When $\pi_n = \pi$ for all n, $\bar{\Phi}$ reduces to the expected potential $\mathbb{E}_{\theta\sim\pi}[\Phi(\mathbf{x}; \theta)]$, recovering the classical potential game structure.*

Proposition 2 implies that any variational generalized Nash equilibrium (v-GNE) of the belief-dependent game $\mathcal{G}(\boldsymbol{\pi})$ corresponds to a stationary point of the constrained optimization problem

$$\max_{\mathbf{x}\in\mathcal{X}_c} \bar{\Phi}(\mathbf{x}; \boldsymbol{\pi}) = \max_{\mathbf{x}\in\mathcal{X}_c} \sum_{n\in\mathcal{N}} \mathbb{E}_{\theta\sim\pi_n}\big[f_n(x_n, \mathbf{x}_{-n}; \theta)\big]. \tag{10}$$

Equivalently, v-GNEs satisfy the first-order optimality conditions of (10) under standard constraint qualifications, providing a convenient variational characterization for decentralized learning and analysis.

3 BELIEF AND LEARNING MODEL

The game $\mathcal{G}(\theta)$ depends on an unknown parameter θ. The induced interaction is therefore a GNG with incomplete information, where agents must act based on beliefs inferred from observations. To address this uncertainty, we introduce a probabilistic observation model that induces a stochastic sequence of belief-dependent GNGs.

Assumption 3 (Communication graph). *The communication graph $G = (\mathcal{N}, E)$ is fixed, undirected, and connected. Agents may become inactive over time; the active agent set is denoted by $\mathcal{N}_n^{(t)} \subseteq \mathcal{N}$. The active neighborhood of agent n at time t is*

$$\mathcal{V}_n^{(t)} = \{n' \in \mathcal{N}_n^{(t)} : \{n, n'\} \in E\}.$$

Inactive agents keep their last action and do not transmit observations or messages.

Remark 2. *A fixed graph models persistent communication links, while time variation arises through the active set $\mathcal{N}_n^{(t)}$ rather than through topology changes.*

We denote by $\mathcal{I}^{(t)}$ the global information available in the system at time t, and by $\mathcal{I}_n^{(t)}$ the information available to agent n at t.

Assumption 4 (Conditional independence of observations). *Conditioned on $(\mathbf{x}^{(t)}, \theta)$, observations and messages are independent across agents and time. In particular,*

$$P(\mathcal{I}^{(t)} | \mathbf{x}^{(t)}, \theta) = \prod_{n \in \mathcal{N}_n^{(t)}} P_n(\mathcal{I}_n^{(t)} | \mathbf{x}^{(t)}, \theta).$$

This assumption is realistic when observations originate from local sensors or communication channels affected by independent noise, which is common in distributed systems. It is also consistent when, conditional on the global system state, each agent collects data without direct stochastic interaction with others. Finally, it is a standard approximation that enables tractable aggregation laws while remaining close to real decentralized architectures.

3.1 Observation and Information Structure

Let $\theta \in \Theta$ denote the fixed but unknown environment parameter. At each time t, the environment generates signals according to

$$y^{(t)} \sim P(\cdot \mid \mathbf{x}^{(t)}, \theta),$$

where P is a known likelihood model.

Each agent n observes a private signal

$$o_n^{(t)} \sim P_n(\cdot \mid \mathbf{x}^{(t)}, \theta),$$

and receives messages from active neighbors

$$m_{n' \to n}^{(t)} \sim Q_{n',n}(\cdot \mid \mathbf{x}^{(t)}, \theta).$$

The information available to agent n is

$$\mathcal{I}_n^{(t)} = \big(o_n^{(t)}, \{m_{n' \to n}^{(t)} : n' \in \mathcal{V}_n^{(t)}\}\big).$$

3.2 Belief State and Induced Belief-Dependent Game

Each agent maintains a posterior belief

$$\pi_n^{(t)}(\theta) = \mathbb{P}(\theta \mid \mathcal{H}_n^{(t)}), \qquad \mathcal{H}_n^{(t)} = (\mathcal{I}_n^{(0)}, \dots, \mathcal{I}_n^{(t)}).$$

The belief acts as an information state summarizing past observations and induces a sequence of belief-dependent GNGs with expected utilities

$$\bar{f}_n(x_n, \mathbf{x}_{-n}; \pi_n^{(t)}) = \mathbb{E}_{\theta \sim \pi_n^{(t)}}\big[f_n(x_n, \mathbf{x}_{-n}; \theta, \mu)\big].$$

Remark 3 (Policy regularity). *We assume that each agent's action selection rule is measurable and depends only on the current action profile and local belief, i.e., agent n's policy satisfies*

$$x_n^{(t+1)} \sim \kappa_n(\cdot \mid \mathbf{x}^{(t)}, \pi_n^{(t)}),$$

for some measurable stochastic kernel κ_n. In particular, the posterior-sampling best-response policy defined below satisfies this property.

Proposition 3 (Markov belief state). *The augmented local state*

$$s_n^{(t)} = (\mathbf{x}^{(t)}, \pi_n^{(t)})$$

is Markov for agent n.

Proof. By Bayes' rule, the posterior $\pi_n^{(t)}$ is a sufficient statistic of the local information history $\mathcal{H}_n^{(t)}$ for inference about θ.

Under Assumption 4, the observation model satisfies

$$P(\mathcal{I}_n^{(t+1)} \mid \theta, \mathbf{x}^{(0:t+1)}, \mathcal{H}_n^{(t)}) = P(\mathcal{I}_n^{(t+1)} \mid \theta, \mathbf{x}^{(t+1)}),$$

i.e., future observations are conditionally independent of past observations given $(\mathbf{x}^{(t+1)}, \theta)$. The action update rule is measurable with respect to $(\mathbf{x}^{(t)}, \pi_n^{(t)})$. Therefore, the conditional distribution of $(\mathbf{x}^{(t+1)}, \pi_n^{(t+1)})$ depends only on $(\mathbf{x}^{(t)}, \pi_n^{(t)})$. Hence, $s_n^{(t)}$ is a Markov state. □

Beliefs evolve recursively according to Bayes' rule

$$\pi_n^{(t+1)}(\theta) \propto P_n(o_n^{(t+1)} \mid \mathbf{x}^{(t+1)}, \theta) \prod_{n' \in \mathcal{V}_n^{(t+1)}} Q_{n',n}(m_{n' \to n}^{(t+1)} \mid \mathbf{x}^{(t+1)}, \theta)\, \pi_n^{(t)}(\theta).$$

3.3 Belief-Based Best-Response Dynamics

At time t, agents compute best-responses in the belief-dependent game. In practice, we consider a posterior-sampling implementation where agent n samples

$$\tilde{\theta}_n^{(t)} \sim \pi_n^{(t)},$$

and computes $x_n^{(t+1)} \in \arg\max_{x_n \in X_n(\mathbf{x}_{-n}^{(t)})} f_n(x_n, \mathbf{x}_{-n}^{(t)}; \tilde{\theta}_n^{(t)})$.

This defines a decentralized posterior-sampling policy. The resulting stochastic process $(\mathbf{x}^{(t)}, \boldsymbol{\pi}^{(t)})_{t \geq 0}$ can therefore be interpreted as decentralized stochastic best-response dynamics evolving across a sequence of belief-dependent generalized Nash games.

4 PENALIZED LEARNING DYNAMICS

We consider decentralized learning in belief-dependent generalized Nash games with coupled constraints. Each agent combines posterior sampling with a quadratic penalty to enforce constraints, enabling fully local updates without coordination of dual variables.

4.1 Penalized Payoff and Action Update

Agent n maximizes a penalized objective

$$J_n(x_n, \mathbf{x}_{-n}; \tilde{\theta}_n^{(t)}, \lambda^{(t)}) = f_n(x_n, \mathbf{x}_{-n}; \tilde{\theta}_n^{(t)}) - \sum_{n' \in \mathcal{V}_n} \lambda_{n,n'}^{(t)}\, g_{n,n'}(x_n, x_{n'})^2, \tag{11}$$

where $\tilde{\theta}_n^{(t)} \sim \pi_n^{(t)}$ is a posterior sample and $\lambda_{n,n'}^{(t)} \geq 0$ is the local penalty.

Actions are updated via projected gradient ascent:

$$x_n^{(t+1)} = \mathrm{Proj}_{\mathcal{X}_n}\Big[x_n^{(t)} + \alpha^{(t)} \nabla_{x_n} J_n(x_n^{(t)}, \mathbf{x}_{-n}^{(t)}, \tilde{\theta}_n^{(t)}, \lambda^{(t)})\Big],$$

with step sizes $(\alpha^{(t)})$ satisfying standard stochastic approximation conditions.

4.2 Penalty Update

Each edge $\{n, n'\}$ updates its penalty proportionally to the constraint violation

$$\lambda_{n,n'}^{(t+1)} = \big[\lambda_{n,n'}^{(t)} + \beta^{(t)} g_{n,n'}(x_n^{(t)}, x_{n'}^{(t)})\big]_+,$$

where $[\,\cdot\,]_+ := \max\{0, \cdot\}$, $(\beta^{(t)})$ is nonnegative, vanishing, and summable to infinity.

Assumption 5 (Step sizes and penalty growth). *Step sizes $(\alpha^{(t)})$ satisfy $\sum_t \alpha^{(t)} = \infty$, $\sum_t (\alpha^{(t)})^2 < \infty$, and penalties $(\beta^{(t)})$ satisfy $\beta^{(t)} \to 0$, $\sum_t \beta^{(t)} = \infty$.*

No explicit time-scale separation between $(\alpha^{(t)})$ and $(\beta^{(t)})$ is assumed.

4.3 Belief Sampling

Each agent samples its environment parameter from the local posterior:

$$\tilde{\theta}_n^{(t)} \sim \pi_n^{(t)},$$

and updates the belief via Bayesian recursion as described in Section 3.

4.4 Distributed Learning Process

The overall dynamics is summarized as follows:

Algorithm 1 Penalized Distributed Thompson Learning

```
for t = 0, 1, 2, ... do
    for each active agent n ∈ N_n^(t) do
        Sample θ̃_n^(t) ~ π_n^(t)
        Update action x_n^(t+1) via projected gradient ascent
        Update local belief π_n^(t+1) via Bayes' rule
    for each edge {n, n'} ∈ E do
        Update penalty λ_{n,n'}^(t+1)
```

Communication. Agents exchange information only with neighbors $\mathcal{V}_n$, yielding a per-agent cost $O(|\mathcal{V}_n|)$ independent of total system size. All computations (actions, beliefs, penalties) are local, ensuring scalability and full decentralization.

The process defines a stochastic approximation in the joint space $(\mathbf{x}^{(t)}, \boldsymbol{\pi}^{(t)}, \lambda^{(t)})$.

5 THEORETICAL GUARANTEES AND COMPLEXITY

We provide feasibility, equilibrium, and finite-time guarantees for the penalized distributed learning dynamics. All results assume a connected graph (Assumption 3), independent observations (Assumption 4), regular payoffs (Assumption 1), constraint qualification (Assumption 2), and stochastic approximation with step sizes and diverging penalties (Assumption 5).

When relevant, we exploit the potential-game structure from Section 2.

Let $K := \dim(\theta)$ be the dimension of the unknown environment parameter.

5.1 Asymptotic Feasibility

Constraint satisfaction is enforced via the penalty update.

Assumption 6 (Bounded iterates). *$(\mathbf{x}^{(t)})_{t\geq 0}$ remains in a compact subset of $\mathcal{X}$ a.s.*

Theorem 1 (Asymptotic feasibility). *Under Assumptions 6 and 5, every limit point of $(\mathbf{x}^{(t)})$ satisfies the coupled constraints:*

$$g_{n,n'}(x_n, x_{n'}) \leq 0, \quad \forall \{n, n'\} \in E.$$

Sketch. The quadratic penalty acts as stochastic dual ascent. Since $\sum_t \beta^{(t)} = \infty$ and $\beta^{(t)} \to 0$, persistent violations would make multipliers diverge, contradicting stability under bounded iterates and Lipschitz gradients. Hence, violations vanish asymptotically. □

5.2 Convergence to Generalized Nash Equilibrium

We now incorporate learning and equilibrium structure.

Assumption 7 (Posterior consistency). *For each agent n, the local belief converges almost surely to the true parameter:*

$$\pi_n^{(t)} \Rightarrow \delta_\theta.$$

Assumption 8 (Potential structure). *For every $\theta \in \Theta$, the full-information game $\mathcal{G}(\theta)$ is an exact potential game with potential $\Phi(\cdot; \theta)$ (Section 2).*

Theorem 2 (Almost sure convergence to GNE). *Under Assumptions 1–2, 5, 6, 7, and 8, the iterates $(\mathbf{x}^{(t)})$ converge almost surely to the set of variational generalized Nash equilibria of $\mathcal{G}(\theta)$.*

Proof. Posterior consistency ensures $\tilde{\theta}_n^{(t)} \to \theta$ a.s., so the stochastic gradient decomposes as

$$\nabla J_n(\mathbf{x}^{(t)}; \tilde{\theta}_n^{(t)}) = \nabla J_n(\mathbf{x}^{(t)}; \theta) + \varepsilon_n^{(t)}, \quad \varepsilon_n^{(t)} \to 0 \text{ a.s.}$$

Hence the dynamics is a stochastic approximation with vanishing noise. Standard Robbins–Monro results imply convergence to stationary points of the true penalized potential. By asymptotic feasibility and the potential property, these coincide with variational GNE of $\mathcal{G}(\theta)$. □

5.3 Finite-Time Bayesian Regret

We quantify the cost of learning through Bayesian regret.

Definition 6 (Bayesian regret). *The regret of agent n over horizon T is*

$$R_n(T) = \sum_{t=0}^{T-1} \Big(V_n^{\pi^*}(s^{(t)}; \theta) - \mathbb{E}[V_n^{\pi^{(t)}}(s^{(t)}; \theta)]\Big).$$

Assumption 9 (Linear value representation). *There exists a bounded feature map ϕ_n such that*

$$V_n^{\pi}(s; \theta) = \langle \phi_n^{\pi}(s), \theta \rangle.$$

Theorem 3 (Bayesian regret bound). *Under Assumption 9,*

$$\mathbb{E}[R_n(T)] \le \frac{M}{1-\gamma}\sqrt{2T\, I(\theta; H_T)}.$$

Proof. The result follows from the information-theoretic analysis of posterior sampling. The belief-dependent game induces a bandit over policies whose reward is Lipschitz in θ. Bounding the mutual information yields sublinear regret. □

5.4 Computational Complexity and Scalability

We analyze the per-iteration complexity of the algorithm.

Proposition 4 (Per-iteration complexity). *Each agent update requires*

$$O(|\mathcal{V}_n| + K)$$

operations, where $|\mathcal{V}_n|$ is the neighborhood size and $K = \dim(\theta)$ is the latent parameter dimension.

Proof. Gradient evaluation depends only on neighbors' actions and local belief statistics. Posterior sampling and belief updates scale linearly in K, while penalty updates scale linearly in $|\mathcal{V}_n|$. □

Proposition 5 (Communication complexity). *The communication load per iteration is $O(|E|)$ messages and does not scale with the number of observations.*

Compared to centralized equilibrium computation requiring global optimization over $\mathcal{X}_c$, the proposed dynamics scales linearly with network size and remains fully decentralized.

These results establish convergence of decentralized learning towards belief-consistent GNEs with provable finite-time efficiency and scalable implementation.

6 EFFICIENCY LOSS UNDER DECENTRALIZED INFORMATION

The previous section establishes that the learning dynamics converges towards a GNE of the true game $\mathcal{G}(\theta)$. We now quantify the efficiency loss induced solely by decentralized information.

Throughout this section, expectations are taken with respect to the prior on θ and all sources of randomness in the learning dynamics.

6.1 Centralized and Decentralized Benchmarks

Let the social welfare be defined as

$$W(\mathbf{x}; \theta) := \sum_{n \in \mathcal{N}} f_n(x_n, \mathbf{x}_{-n}; \theta, \mu). \tag{12}$$

We compare two information structures.

Centralized information. Agents observe the true parameter θ and can condition their actions on it.

Decentralized information. Each agent n only observes its local information filtration and forms beliefs $\pi_n^{(t)}$ according to the learning dynamics described previously.

Assumption 10 (Well-posed social welfare maximization). *For every θ:*

(1) *Action sets $\mathcal{X}_n$ are compact and convex.*
(2) *$f_n(\cdot, \mathbf{x}_{-n}; \theta)$ is continuous in $\mathbf{x}$.*
(3) *$W(\mathbf{x}; \theta)$ is bounded.*

This guarantees existence of maximizers and finiteness of expectations.

Policy classes. Let Π^{full} denote the set of admissible policies measurable with respect to full information $(\theta, \text{history})$.

Let Π^{dec} denote the set of admissible policies measurable with respect to local information and local beliefs $(\pi_n^{(t)}, \mathcal{I}_n^{(t)})$.

Define

$$W^{\text{full}} := \sup_{\pi \in \Pi^{\text{full}}} \mathbb{E}[W(\mathbf{x}_\pi; \theta)], \tag{13}$$

$$W^{\text{dec}} := \sup_{\pi \in \Pi^{\text{dec}}} \mathbb{E}[W(\mathbf{x}_\pi; \theta)]. \tag{14}$$

By construction,

$$W^{\text{dec}} \le W^{\text{full}}. \tag{15}$$

6.2 Price of Miscoordination

Definition 7 (Price of Miscoordination).

$$\text{PoM} := \frac{W^{\text{full}}}{W^{\text{dec}}} \ge 1. \tag{16}$$

This quantity isolates the efficiency loss generated purely by informational decentralization, holding incentives and feasibility constraints fixed.

6.3 Connection with Equilibrium Structure

We now relate decentralized social welfare to equilibrium outcomes of the decentralized game.

Assumption 11 (Potential structure). *There exists a potential function $\Phi(\mathbf{x}; \theta)$ such that*

$$\nabla_{x_n} f_n(\mathbf{x}; \theta) = \nabla_{x_n} \Phi(\mathbf{x}; \theta), \quad \forall n. \tag{17}$$

Assumption 12 (Equilibrium efficiency within information class). *Among all decentralized policies, the social welfare-maximizing ones induce equilibrium play (v-GNE) of the decentralized game.*

This assumption is standard in potential games where equilibria solve a constrained potential maximization problem.

6.4 Relation to Learning Dynamics

Let $\mathbf{x}^{(t)}$ denote the trajectory generated by the decentralized learning algorithm.

Theorem 4 (Asymptotic decentralized efficiency). *Suppose:*

- *the convergence theorem holds,*
- *Assumptions 10–12 hold.*

Then

$$\lim_{t \to \infty} \mathbb{E}[W(\mathbf{x}^{(t)}; \theta)] = W^{\text{dec}}. \tag{18}$$

Proof. By the convergence theorem, $\mathbf{x}^{(t)}$ converges (in probability or almost surely depending on the result used) to the set of GNEs of the decentralized game, including v-GNEs.

Under the potential structure, v-GNEq correspond to stationary points (i.e., optima) of the decentralized potential maximization problem.

Under Assumption 12, social welfare-maximizing decentralized policies induce equilibrium play (v-GNEs). Therefore, the limiting social welfare equals the optimal social welfare achievable under decentralized information. The theorem statement follows. □

6.5 Finite-Time Welfare Loss

Let $R_n(T)$ denote the cumulative regret of agent n.

Theorem 5 (Finite-time inefficiency bound). *Under bounded rewards and sublinear regret for each agent,*

$$W^{\mathrm{dec}} - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W(\mathbf{x}^{(t)};\theta)] \le \sum_{n\in\mathcal{N}}\frac{R_n(T)}{T}. \tag{19}$$

Proof. Define the decentralized optimal trajectory $\mathbf{x}^{\star,\mathrm{dec}}$.

By additivity of social welfare,

$$W(\mathbf{x}^{\star,\mathrm{dec}};\theta) - W(\mathbf{x}^{(t)};\theta) = \sum_{n}\left(f_n(x_n^{\star,\mathrm{dec}}, \mathbf{x}_{-n}^{\star,\mathrm{dec}};\theta) - f_n(x_n^{(t)}, \mathbf{x}_{-n}^{(t)};\theta)\right).$$

Summing over t and taking expectations yields

$$TW^{\mathrm{dec}} - \sum_{t=1}^{T}\mathbb{E}[W(\mathbf{x}^{(t)};\theta)] \le \sum_{n} R_n(T).$$

Divide by T to conclude. □

6.6 Structural vs Learning Loss

The Price of Miscoordination decomposes total inefficiency into

(1) **Structural loss**

$$W^{\mathrm{full}} - W^{\mathrm{dec}},$$

caused by informational decentralization.

(2) **Learning loss**

$$W^{\mathrm{dec}} - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W(\mathbf{x}^{(t)};\theta)],$$

which vanishes as regret becomes sublinear.

If $R_n(T) = O(\sqrt{T})$, the learning loss decays at rate $O(T^{-1/2})$.

7 NUMERICAL ILLUSTRATION

We illustrate our decentralized learning framework in a spatial coordination problem inspired by asymmetric counterinsurgency (COIN) scenarios. Blue agents represent coordinated units (e.g., military forces), red agents are insurgents clustered in strategic strongholds (*bastions*), and civilians provide partial, local information. The connection with classical combat models (Lanchester [11, 13]) motivates modeling interactions between opposing forces, extending attrition-based dynamics to *spatial positioning and information-driven engagement.*

7.1 Setting

Agents evolve in a 25×25 domain under a fixed connected communication graph. The latent parameter

$$\theta^{\star} = (12, 15, 21, 13, 9), \quad K = 5$$

specifies red-agent allocation across K bastions. Red agents are sampled from Gaussians around bastions (σ_r), while civilian agents move stochastically near bastions (σ_c), generating local observations communicated to nearby blue agents under limited communication.

The environment contains 95 blue, 70 red, and $\{115, 145, 185\}$ civilian agents, with civilian distribution proportional to $\theta^{\star}$, correlating information availability with insurgent presence. The time horizon is $T = 60$, sensing radii are $\rho = 1.3$ (blue) and $\bar{\rho} = 1.5$ (civilians), and trust levels $\mu \in \{0, 0.65, 1\}$ control the reliability of civilian-provided information. Blue agents optimize identical utilities; only the information structure varies, isolating the value-of-information effect.

Each curve averages 50 independent runs; full parameters are in Table 1.

7.2 Convergence and Feasibility

Constraint satisfaction and equilibrium convergence (Theorems 1–2) are measured via

$$\mathrm{Viol}(t) = \frac{1}{|E|}\sum_{\{n,n'\}\in E}[g_{n,n'}(x_n^{(t)}, x_{n'}^{(t)})]_+, \quad G(t) = \|\nabla J(\mathbf{x}^{(t)};\theta)\|.$$

The pseudo-gradient norm $G(t)$ vanishes while joint value stabilizes, indicating convergence to a stationary GNE, empirically reached after ~ 38 iterations.

7.3 Finite-Time Regret

Empirical social welfare regret is

$$\widehat{R}(T) = \sum_{t=0}^{T-1}\left(\widehat{V}^{*} - W(\mathbf{x}^{(t)};\theta)\right),$$

with $\widehat{V}^{*}$ estimated from long-run equilibrium. Regret grows sublinearly, $\widehat{R}(T) = O(\sqrt{T})$, matching information-theoretic bounds. Figure 1 compares empirical value trajectories to the theoretical bound, showing early exploration ($t \approx 15$) followed by near-efficient stabilization.

Thompson sampling outperforms UCB and ϵ-greedy, which produce higher cumulative regret due to over- or inefficient exploration.

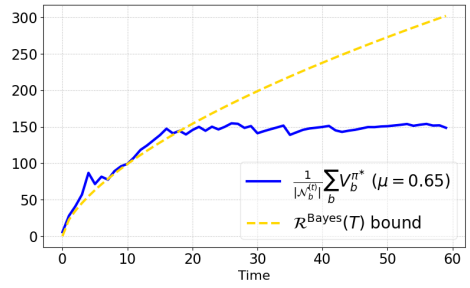


Figure 1: Convergence analysis.

7.4 Efficiency Loss and Value of Information

Average social welfare and empirical Price of Miscoordination are

$$\widehat{W}(T) = \frac{1}{T}\sum_{t=1}^{T} W(\mathbf{x}^{(t)};\theta), \quad \widehat{\mathrm{PoM}} = \frac{\widehat{W}^{\mathrm{full}}}{\widehat{W}(T)}.$$

Figure 2 shows social welfare versus trust μ: performance rises sharply from 0 to 0.65 with diminishing returns beyond. Civilian number has limited impact, highlighting that coordination relies more on reliability than observation volume. Communication-enabled regimes outperform no-communication baselines. After transients, $\widehat{\text{PoM}} > 1$ indicates persistent efficiency loss from decentralized information.

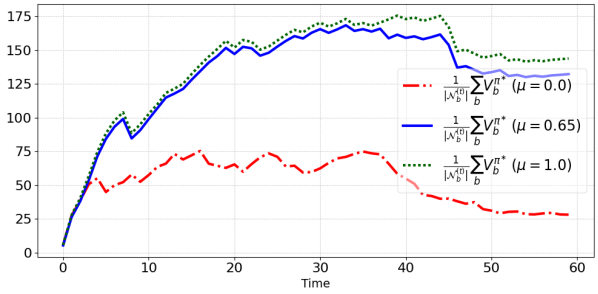


Figure 2: Impact of trust μ on social welfare.

Figure 3 reports $\widehat{\text{PoM}}$ versus sensing radius $\bar{\rho}$: small radii fragment beliefs, larger radii improve coordination until saturation, indicating global belief alignment.

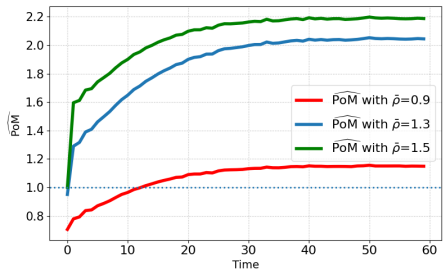


Figure 3: Sensitivity of Empirical Price of Miscoordination to sensing radius.

Spatial configurations (Figs. 5a–5d, Appendix A.4) illustrate decentralized learning: agents initially disperse with misaligned beliefs, then cluster near high-probability regions inferred via shared observations, stabilizing near latent centers, consistent with equilibrium convergence and uncertainty reduction.

7.5 Summary of Empirical Finding

The experiments confirm the theoretical results. Constraint violations vanish and equilibria emerge. Regret is sublinear and remains below classical exploration baselines. Decentralized information induces a bounded efficiency loss. Coordination performance depends primarily on information quality rather than observation quantity. These results validate the learning dynamics and support the interpretation of the Price of Miscoordination as a measurable coordination efficiency gap.

8 CONCLUSION

This framework enables the simulation of interactions between clusters of strategic agents under uncertainty and decentralized information, a setting common in defense and security contexts. By combining Bayesian learning and game-theoretic equilibrium analysis, it captures both adversarial adaptation and cooperative coordination. The approach supports realistic modeling of multi-force interactions and information-driven strategic behavior.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47 (2002), 235–256.

[2] Seymour J Deitchman. 1962. A Lanchester model of guerrilla warfare. *Operations Research* 10, 6 (1962), 818–827.

[3] Zhenhua Deng and Tao Chen. 2024. Distributed Nash equilibrium seeking for constrained multicluster games of second-order nonlinear multiagent systems. *IEEE Trans. Automat. Control* 69, 11 (2024), 7855–7862.

[4] Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. 2023. Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*. PMLR, 9343–9373.

[5] Michael I Jordan, Tianyi Lin, and Manolis Zampetakis. 2023. First-order algorithms for nonlinear generalized Nash equilibrium problems. *Journal of Machine Learning Research* 24, 38 (2023), 1–46.

[6] Edward H. Kaplan, Moshe Kress, and Roberto Szechtman. 2010. Confronting Entrenched Insurgents. *Operations Research* 58, 2 (2010), 329–341.

[7] Amin Karbasi, Nikki Lijing Kuang, Yian Ma, and Siddharth Mitra. 2023. Langevin thompson sampling with logarithmic communication: bandits and reinforcement learning. In *International Conference on Machine Learning*. PMLR, 15828–15860.

[8] Moshe Kress and Niall J MacKay. 2014. Bits or shots in combat? The generalized Deitchman model of guerrilla warfare. *Operations Research Letters* 42, 1 (2014), 102–108.

[9] Moshe Kress and Roberto Szechtman. 2009. Why defeating insurgencies is hard: The effect of intelligence in counterinsurgency operations—A best-case scenario. *Operations Research* 57, 3 (2009), 578–585.

[10] Ashok K. S. Krishnan, Hélène Le Cadre, and Ana Busic. 2025. How Irrationality Shapes Nash Equilibria: A Prospect-Theoretic Perspective. In *Proceedings of the 64th International Conference on Decision and Control*. IEEE, Rio de Janeiro, Brazil.

[11] F. W. Lanchester. 1916. Aircraft in warfare: the dawn of the aerial age. *Electrical World and Engineer* 59 (1916), 590–592.

[12] Frederick William Lanchester. 1916. *Aircraft in warfare: The dawn of the fourth arm.* Constable limited.

[13] F. W. Lanchester. 1916. Mathematics of warfare. *Electrical World and Engineer* 59 (1916), 589–591.

[14] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[15] Ian Osband, Daniel Russo, and Benjamin Van Roy. 2013. More efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[16] Shuang Qiu, Ziyu Dai, Han Zhong, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. 2023. Posterior sampling for competitive rl: Function approximation and partial observation. *Advances in neural information processing systems* 36 (2023), 26585–26637.

[17] Daniel Russo and Benjamin Van Roy. 2016. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research* 17, 68 (2016), 1–30.

[18] David et al. Silver. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489. https://doi.org/10.1038/nature16961

[19] William R. Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[20] Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Ménard. 2023. Model-free posterior sampling via learning rate randomization. *Advances in Neural Information Processing Systems* 36 (2023), 73719–73774.

[21] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. 2019. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv preprint arXiv:1911.10635* (2019).

[22] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *Handbook of Reinforcement Learning and Control* (2021), 321–384.

A SIMULATION DETAILS AND ADDITIONAL RESULTS

This appendix provides detailed parameters and visualizations supporting the numerical experiments presented in the main text. All results illustrate decentralized learning dynamics, constraint satisfaction, and spatial coordination in a COIN-inspired scenario.

A.1 Experimental Parameters

Table 1 summarizes all simulation parameters, including agent counts, sensing radii, mobility, and latent environment settings. These values reflect a realistic asymmetric setting with blue forces, red insurgents, and civilian informants.

Variable	Description	Value
$\mathcal{N}_B^{(0)}$	Number of blue agents (soldiers)	95
$\mathcal{N}_C$	Number of civilian agents	115, 145, 185
$\mathcal{N}_R^{(0)}$	Number of red agents (insurgents)	70
K	Number of strongholds	5
Z_1–Z_5	Coordinates of strongholds	See Fig. 5
T	Time horizon	60
ρ	Blue agents sensing radius	1.3
$\bar{\rho}$	Civilian sensing / comm. radius	1.5
σ_r	Insurgents dispersion	1.1
σ_c	Civilian mobility	2
$\beta_0, \beta_1, \beta_2$	Engagement score parameters	3, 0.2, 4.0
μ	Trust parameter	0.0, 0.65, 1.0
α	Utility trade-off	1.5
$\theta^\star$	Latent allocation of insurgents	(12, 15, 21, 13, 9)

Table 1: Simulation parameters for blue agents, red agents, civilians, and environment.

A.2 Utility Design under Uncertainty and Decentralized Information

The payoff of each blue agent b is constructed to capture the trade-off between collective engagement effectiveness and individual exposure to risk in a decentralized and uncertain environment. Formally, the expected utility of agent b is defined as the difference between a collective utility term U_b and a subjective cost C_b, as defined in Equation (1). Further, we define a belief-weighted threat function

$$p(\mu, \hat{\theta}_b^{(t)}, \psi(x_r)) = \mu + (1-\mu)\,\frac{\langle \hat{\theta}_b^{(t)}, \psi(x_r)\rangle}{n_r},$$

which combines prior trust in civilian information with belief-based evidence on insurgent presence. Here, $\hat{\theta}_b^{(t)}$ denotes agent b's posterior belief over strongholds, while $\psi(x_r)$ encodes the location of insurgent r. This quantity captures how agents balance exogenous information reliability and endogenous inference: large μ induces conservative behavior, whereas smaller μ increases reliance on local beliefs. The perceived threat directly enters the engagement score

$$z_{b,r} = \beta_0 - \beta_1 \|x_b - x_r\|^2 + \beta_2\, p(\mu, \hat{\theta}_b^{(t)}, \psi(x_r)),$$

which is mapped to a probability of successful neutralization via the logistic function

$$\sigma(z_{b,r}) = \frac{1}{1+\exp(-z_{b,r})}.$$

The collective component relies on a probabilistic engagement model: for each observed insurgent r, agent b forms an engagement score

$$z_{b,r} = \beta_0 - \beta_1 \|x_b - x_r\|^2 + \beta_2\, p(\mu, \hat{\theta}_b, \psi(x_r)),$$

which combines a distance-based attenuation term with a belief-weighted threat level. The latter depends on the agent's posterior estimate $\hat{\theta}_b$ of insurgent allocation across strongholds and is modulated by a trust parameter μ, balancing civilian information and local evidence. The score is mapped through a logistic function $\sigma(z_{b,r})$ to obtain a smooth probability of successful neutralization.

Assuming independent engagement attempts, the probability that an insurgent is neutralized by the neighborhood $\mathcal{V}_b$ is expressed as a complement of a product of failure probabilities, yielding a differentiable approximation of collective success. The collective utility U_b is then defined as the sum of these probabilities over observed insurgents, capturing the agent's marginal contribution to team performance.

In parallel, the individual cost C_b models vulnerability as a function of proximity to insurgents, given by a sum of squared distances augmented by a stochastic perturbation capturing environmental uncertainty. This construction yields a smooth, spatially grounded payoff function that is compatible with gradient-based decentralized optimization, while faithfully encoding the interaction between belief-driven engagement and risk-aware positioning.

A.3 Constraint Satisfaction and Convergence

Figures 4a–4b illustrate the evolution of feasibility and equilibrium convergence. Constraint violations are measured by

$$\mathrm{Viol}(t) = \frac{1}{|E|} \sum_{\{n,n'\}\in E} \left[g_{n,n'}(x_n^{(t)}, x_{n'}^{(t)}) \right]_+,$$

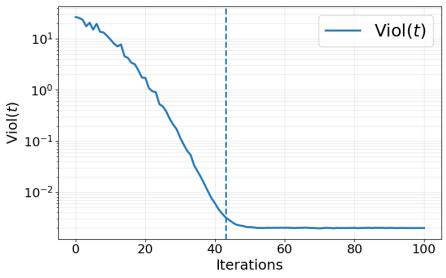
while convergence to a stationary GNE is assessed via the pseudo-gradient norm

$$G(t) = \|\nabla J(\mathbf{x}^{(t)}; \theta)\|.$$

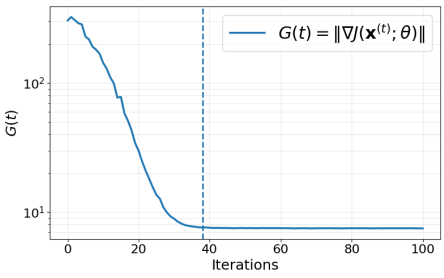
Empirically, violations decay to near-zero and the pseudo-gradient vanishes, confirming convergence predicted by Theorems 1–2.

A.4 Spatial Dynamics

Figures 5a–5d depict representative agent configurations over time. Initially, blue agents are dispersed with misaligned beliefs. As learning progresses, they aggregate near high-probability regions inferred from civilian observations. By the stationary regime, positions stabilize around latent strongholds, reflecting reduced uncertainty and coordinated coverage.

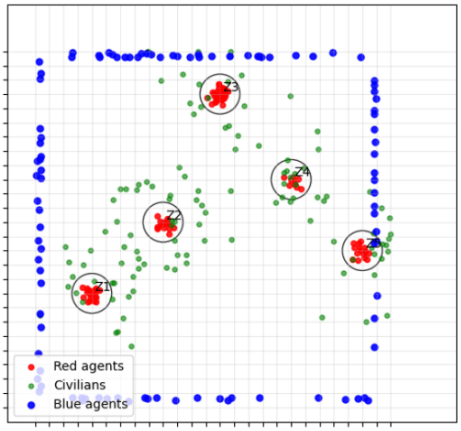


(a) Constraint violation $\mathrm{Viol}(t)$

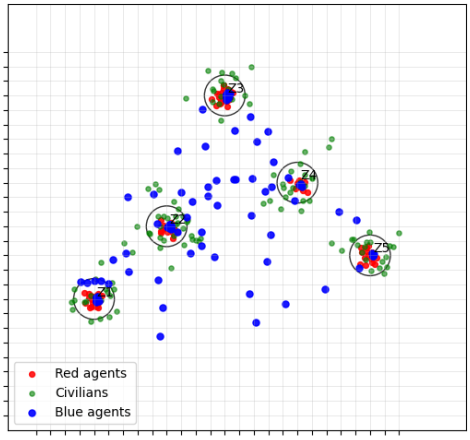


(b) Pseudo-gradient norm $G(t)$

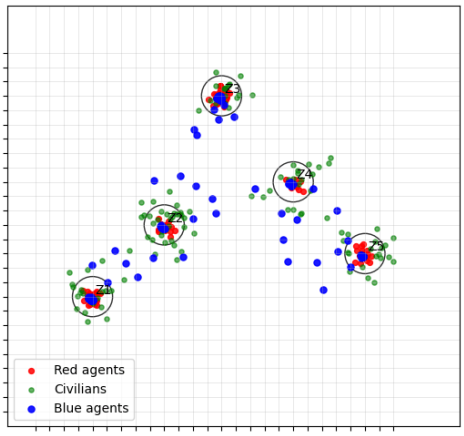
Figure 4: Constraint satisfaction and equilibrium convergence over iterations.



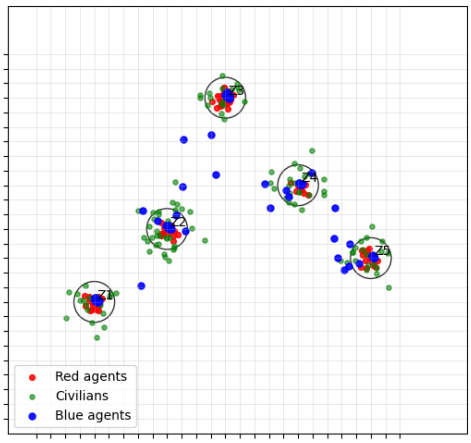
(a) Initial positions



(b) Early learning



(c) Belief alignment



(d) Stationary regime

Figure 5: Evolution of blue-agent positions over time, illustrating decentralized coordination around inferred red-agent strongholds.