

Decentralized Optimal Equilibrium Learning Over Dynamic Networks

Seref Taha Kiremitci¹ and Muhammed O. Sayin¹

Abstract—This paper studies decentralized learning of socially optimal equilibria in finite normal-form games over dynamic communication networks. Each agent observes only its own realized payoffs, does not know the game a priori, and can communicate only with time-varying neighbors using low-bandwidth messages. We propose networked decentralized optimal equilibrium learning dynamics in which agents generate randomized semantic content/discontent signals from local payoff comparisons and exchange time-stamped time-stacked tables rather than raw actions, payoff information or local estimates/parameters. The method combines table fusion with temporal majority reconstruction to mitigate dynamic communication while preserving fully decentralized operation. We establish finite-time logarithmic regret guarantees, with an in-phase exploration perturbation, for optimal equilibrium selection under utilitarian and proportional-fair social welfare objectives. Simulation results further show that the proposed approach can effectively select socially desirable equilibria over dynamic communication networks.

I. INTRODUCTION

Decentralized multi-agent systems are increasingly central to modern engineered platforms, including networked control systems and distributed autonomous decision-making [1]–[3]. In such systems, a critical challenge is the balance between *stability* and *efficiency*. Equilibrium, providing robustness against unilateral deviations, need not be socially desirable. Welfare-maximization, being socially desirable, need not be stable under strategic behavior. Optimal equilibrium selection can balance stability and efficiency. Nevertheless, it is largely unexplored in the context of decentralized learning.

For example, in repeated play of normal-form games, decentralized payoff-based learning dynamics have been developed to promote Pareto-efficient solutions in interdependent games [4], and have been extended to the selection of Pareto-efficient Nash equilibria in distributed systems [5]. Recently, [6] has addressed decentralized learning of equilibrium maximizing the minimum transformed utility. All [4]–[6] rely on stochastic stability guarantees as certain exploration parameters decay to zero. However, lower exploration levels lead to slower learning rates, limiting the practicality of such dynamics. Decentralized optimal equilibrium learning has also been studied in the context of stochastic games yet only for special classes of games, such as common-interest games and no-conflict games [7]–[9]. All these works do not use any information exchange across agents while communication can play a crucial role in effective coordination [10], [11].

¹Seref Taha Kiremitci and Muhammed O. Sayin are with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Türkiye. Emails: taha.kiremitci@bilkent.edu.tr, sayin@ee.bilkent.edu.tr

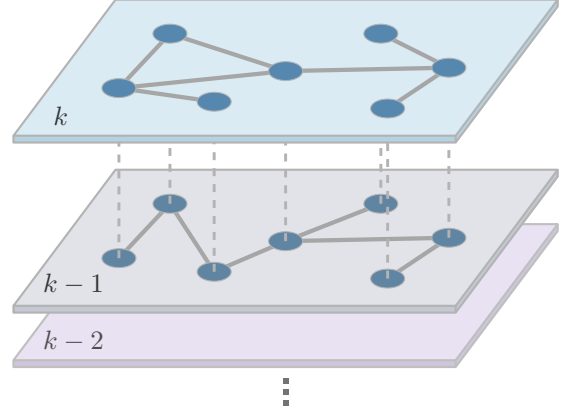


Fig. 1: An illustration of decentralized learning over dynamic communication networks. Layers depict time-varying topology at instances $k - 2, k - 1, k$.

Rather recently, [12] showed that single-bit communication can be effective to drive broad classes of general games toward socially optimal outcomes, while [13] extended this idea to decentralized optimal equilibrium learning in stochastic games under general social welfare objectives. Both [12] and [13] use probabilistic *content/discontent signals* rather than direct sharing of model parameters or estimates, and they provide explicit logarithmic finite-time guarantees. However, both of these schemes rely on fully connected and reliable communication, which is restrictive, e.g., in geographically distributed networks, and can make the learning mechanism sensitive to temporal link failures.

This paper addresses decentralized optimal equilibrium learning in repeated play of normal-form games over dynamic and unreliable communication links, as illustrated in Fig. 1. Agents observe only their own realized payoffs, do not know the game a priori, and communicate through finite-support semantic messages. Dynamic networks pose challenges that are not a direct extension of [12], [13]: content/discontent bits are phase-specific, so they cannot be averaged across time as in consensus, and neighbor-only failing links make network-wide bits delayed, missing, and statistically dependent across overlapping table windows.

Relative to [12], [13], which assume fully connected reliable broadcast of single-bit signals, our contributions are: (i) Networked-DOEL, which reconstructs network-wide content from time-stamped, time-stacked tables, conservative fusion, and temporal majority polling; (ii) a finite-time regret bound of the same form as in those works, but for dynamic neighbor-only graphs, by restoring independence of delayed

access events rather than only rescaling success probabilities; (iii) experiments on a restricted Erdős–Rényi backbone with dropouts, including the complete-graph communication setting of [13]. We examine both utilitarian (sum of local payoffs) and proportional-fair (product of local payoffs) welfare objectives.

Our work is related more broadly to the large literature on networked coordination, but differs from it in an essential way. In distributed optimization and consensus over graphs, the network is typically used to aggregate estimates, gradients, or state variables toward a common point. This perspective includes distributed optimization or reinforcement learning over time-varying directed graphs [10], [11], as well as communication-efficient schemes based on sign-based or compressed information exchange [14]–[17]. Such works show that sparse and time-varying network structure can be exploited effectively, but their objectives are fundamentally cooperative and usually rely on structural properties such as convexity and smoothness, specifically for payoff-sum maximization. In contrast, our problem is game-theoretic. Here, the agents are strategically coupled and the target is reaching optimal equilibrium over the highly unstructured equilibrium set with respect to broad welfare objectives beyond payoff-sum maximization even under dynamic communication networks.

The rest of the paper is organized as follows. Section II introduces the problem formulation. Section III presents the proposed networked decentralized optimal equilibrium learning dynamics. Section IV and Section V, respectively, provide analytical and numerical results. Finally, Section VI concludes the paper.

II. PROBLEM FORMULATION

Consider an n -agent *normal-form* game characterized by the tuple $\mathcal{G} := \langle A^i, u^i \rangle_{i \in \mathcal{N}}$, where $\mathcal{N} := \{1, \dots, n\}$ is the index set of agents, A^i is the finite action set of agent i , $A := \prod_{j \in \mathcal{N}} A^j$ is the joint action space, and $u^i : A \rightarrow \mathbb{R}$ is agent i 's payoff function.

Let $a = (a^i)_{i \in \mathcal{N}} \in A$ be the action profile of all agents and $a^{-i} := (a^j)_{j \neq i}$ denote actions of agents other than agent i . Then, an action profile $a \in A$ is a *pure-strategy ϵ -Nash equilibrium* if, for each agent i ,

$$u^i(a^i, a^{-i}) \geq u^i(\tilde{a}^i, a^{-i}) - \epsilon^i, \quad \forall \tilde{a}^i \in A^i, \quad (1)$$

for some $\epsilon^i \geq 0$. Given $\epsilon = (\epsilon^i)_{i=1}^n$, we define ϵ -equilibrium set as

$$A_\epsilon^{\text{eq}} := \left\{ a \in A : u^i(a) \geq \max_{\tilde{a}^i \in A^i} u^i(\tilde{a}^i, a^{-i}) - \epsilon^i, \quad \forall i \right\}. \quad (2)$$

In general, pure-strategy equilibrium may not exist. However, there always exist sufficiently large tolerance levels ϵ such that a pure-strategy ϵ -equilibrium can exist. Furthermore, different tolerance levels can capture heterogeneous sensitivities to suboptimal responses, ranging from full non-cooperativeness as $\epsilon^i \rightarrow 0$ to full cooperativeness as $\epsilon^i \rightarrow \infty$.

Given the payoffs $(u^i)_{i=1}^n$, the social welfare of all agents is given by

$$\text{SW}(a) := \sum_{i=1}^n w^i \cdot g^i(u^i(a)), \quad (3)$$

for positive weights $w^i > 0$ and monotonically increasing transformations $g^i(\cdot)$.

For example, for $w^i = 1$ and $g^i(x) = x$, the social welfare corresponds to the sum of payoffs $\sum_i u^i(a)$, studied in multi-agent systems for Pareto optimal solutions, e.g., see [4], [5], [10], [11]. On the other hand, for $w^i = 1$ and $g^i(x) = \log(x)$, the social welfare corresponds to the product of payoffs $\log \prod_i u^i(a)$, yielding fair solutions across agents [18].

We say that equilibrium $a_* \in A_\epsilon^{\text{eq}}$ is *optimal* over (non-empty) ϵ -equilibrium set $A_\epsilon^{\text{eq}} \neq \emptyset$ provided that

$$a_* \in \arg \max_{a \in A_\epsilon^{\text{eq}}} \{\text{SW}(a)\}. \quad (4)$$

We consider that agents repeatedly play the game $\mathcal{G}$ to learn optimal equilibria through decentralized learning dynamics. They do not know the model and do not observe the actions of the others. They only observe their own payoff with perfect recall. We let agents communicate with certain other agents (e.g., the ones in proximity) over a communication network through finite support messages while the communication links can change or fail dynamically, as illustrated in Fig. 1.

In this paper, our goal is to *design resilient decentralized optimal equilibrium learning algorithms for such dynamic communication networks*. To quantify the effectiveness of the proposed solution, we focus on regret-based analysis. Particularly, given a time budget T , we define the cumulative regret of the agents' joint play a_t at time t relative to the optimal welfare in (4) as

$$R_T := T \max_{a \in A_\epsilon^{\text{eq}}} \{\text{SW}(a)\} - \sum_{t=1}^T \text{SW}(a_t). \quad (5)$$

III. DECENTRALIZED OPTIMAL EQUILIBRIUM LEARNING OVER NETWORKS

In this section, we first describe Decentralized Optimal Equilibrium Learning (DOEL) framework, in which each agent can reliably communicate with all other agents [13]. Although agents only exchange single-bit messages, as described in Subsection III-A, full connectivity poses a challenge for practical applications. For example, there may not be a direct (one-hop) communication link in-between all agents. To mitigate these issues, we introduce Networked-DOEL dynamics, in which agents can temporally communicate with only a subset of agents over a dynamic communication network.

A. DOEL Dynamics

The DOEL dynamics originally address decentralized equilibrium selection for stochastic games through explore-and-commit and online learning schemes. However, due to

the limited space, here, we only focus on the explore-and-commit scheme for the special case of repeated play of normal-form games, in which the underlying environment has a single state and agents have single-stage objectives.

During the exploitation stage, agents aim to coordinate in the best ϵ -equilibrium. To this end, they need to identify (i) whether the explored action profile is ϵ -equilibrium and (ii) whether it leads to the highest social welfare compared to other equilibria.

To address (i), agents can divide the exploration stage into K phases $k = 1, \dots, K$ with fixed lengths of κ instances. At each phase k , each agent i randomly selects an action $a_{(k)}^i$. Then, within the phase $t = (k-1)\kappa + 1, \dots, k\kappa$, the agent plays according to the distribution $\mathcal{E}^i(a_{(k)}^i) : A^i \rightarrow [0, 1]$ defined by

$$\mathcal{E}^i(a_{(k)}^i)(a^i) = \begin{cases} 1 - \frac{|A^i| - 1}{|A^i|} \cdot \varepsilon^i & \text{if } a^i = a_{(k)}^i \\ \frac{1}{|A^i|} \cdot \varepsilon^i & \text{if } a^i \neq a_{(k)}^i \end{cases}, \quad (6)$$

where $\varepsilon^i \in (0, 1)$ is the *in-phase exploration* probability. Correspondingly, $a_t^i \sim \mathcal{E}^i(a_{(k)}^i)$ independently for each $t = (k-1)\kappa + 1, \dots, k\kappa$.

Each agent can estimate the value of local actions by taking average of the payoffs received whenever the associated action taken. For example, we have

$$\hat{u}_{(k)}^i(a^i) = \frac{\sum_{t=(k-1)\kappa+1}^{k\kappa} u_t^i \cdot \mathbb{I}_{\{a^i = a_{(k)}^i\}}}{\sum_{t=(k-1)\kappa+1}^{k\kappa} \mathbb{I}_{\{a^i = a_{(k)}^i\}}} \quad (7)$$

for all a^i . Then, explored $a_{(k)}^i$ is the ϵ -best response with respect to the estimate $\hat{u}_{(k)}^i$ if

$$\hat{u}_{(k)}^i(a_{(k)}^i) \geq \hat{u}_{(k)}^i(a^i) - \epsilon^i \quad \forall a^i \in A^i. \quad (8)$$

Furthermore, the explored action profile $a_{(k)} = (a_{(k)}^i)_{i \in \mathcal{N}}$ is ϵ -equilibrium if (8) holds for all $i \in \mathcal{N}$.

Recall that agents only observe local actions and payoffs. Even though they can check whether the action explored tolerable, they are not aware of whether everyone's explored actions are tolerable. To mitigate this issue, they can broadcast a single-bit tolerable '1' or intolerable '0' signal with each other at the end of the phase. If everyone signals 1, then the agents can conclude that the explored action profile is ϵ -equilibrium based on the local value estimates.

Remark 1. If agents are self-less, i.e., $\epsilon^i \rightarrow \infty$, then the problem reduces to decentralized welfare optimization and they do not need to divide the exploration stage into phases.

To address (ii), agents need to consider the social welfare (3). However, local payoffs are not necessarily aligned with the social welfare. In the DOEL dynamics, agents broadcast a *randomized* single-bit *content* '1' or *discontent* '0' signal with each other if the local action is tolerable.¹ More explicitly, if $\hat{u}_{(k)}^i(a_{(k)}^i) \geq \max_{a^i} \hat{u}_{(k)}^i(a^i) - \epsilon^i$, then agent i broadcasts

$$s_{(k)}^i = \begin{cases} 1 & \text{with probability } \xi^{w^i(C^i - g^i(\hat{u}_{(k)}^i(a_{(k)}^i)))} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

¹In the payoff-based dynamics of [4], [5], agents locally keep and update a *content/discontent* state without sharing it with the other agents.

independently with some $\xi \in (0, 1)$ and $C^i \geq g^i(u^i(a))$ for all $a \in A$, ensuring that $C^i - g^i(\hat{u}_{(k)}^i(a_{(k)}^i)) \geq 0$ and $\xi^{w^i(C^i - g^i(\hat{u}_{(k)}^i(a_{(k)}^i)))} \in (0, 1]$ is a valid probability. Due to the independent randomization, all agents would broadcast content signals for ϵ -equilibrium $a_{(k)}$ based on the local value estimates with the conditional probability

$$\Pr(s_{(k)}^i = 1 \mid a_{(k)}) = \xi^{C - \text{SW}(a_{(k)})}, \quad (10)$$

where $C = \sum_i w^i C^i$ and $C \geq \text{SW}(a_{(k)})$ by its definition.

Observe that $a_{(k)}$ maximizing the probability (10) maximizes $\text{SW}(\cdot)$ since $\xi \in (0, 1)$. Indeed, the distribution over action profiles conditioned on content-endorsement by all agents corresponds to the entropy-regularized welfare maximization over the equilibrium set [13, Remark 2].

Another key challenge is that agents do not observe the explored action profile. Relying only on local observations and communication, the DOEL dynamics let agents commit to the local action that is the most frequently content-endorsed by all agents. Under mild assumptions, [13] showed that the DOEL dynamics can attain at most logarithmic regret up to an in-phase exploration offset.

B. Networked-DOEL Dynamics

In the Networked-DOEL dynamics, agents are not able to reliably broadcast signals to all other agents. The communication network might dynamically change across phases, e.g., see Fig. 1. To resolve this issue, the Networked-DOEL dynamics let agents share time-stamped and time-stacked messages, as illustrated in Fig. 2.

Time-Stamped Signals: Content/discontent signals are specific to the current actions explored. However, the actions independently got explored according to the uniform distribution across phases. Therefore, agents are not able to diffuse the content signals across the underlying communication network by aggregating all information shared as in consensus networks. Instead, they can share the content/discontent signals with a timestamp to identify the phases associated with the signals. They not only share their own content/discontent signals but also share the most recent content/discontent signals of all others based on the received signals from the neighboring agents, as commonly used for dynamic environments, e.g., see [19].

Time-Stacked Signals: Here all signals matter beyond the most fresh ones. If the signals of agents do not reach to each other, then these agents cannot track the network-wide signals and this can degrade the performance. To mitigate this issue, agents can introduce redundancy by sharing not only the most recent content/discontent signals but also older signals.

More explicitly, let $\hat{s}_{(k),(l)}^{i,j} \in \{0, 1, \emptyset\}$ denote agent i 's belief at phase k about agent j 's content/discontent signal at phase l . Here, $\emptyset$ refers to missing or inconsistent information. Then, at phase k , agent i shares an $n \times m$ dimensional

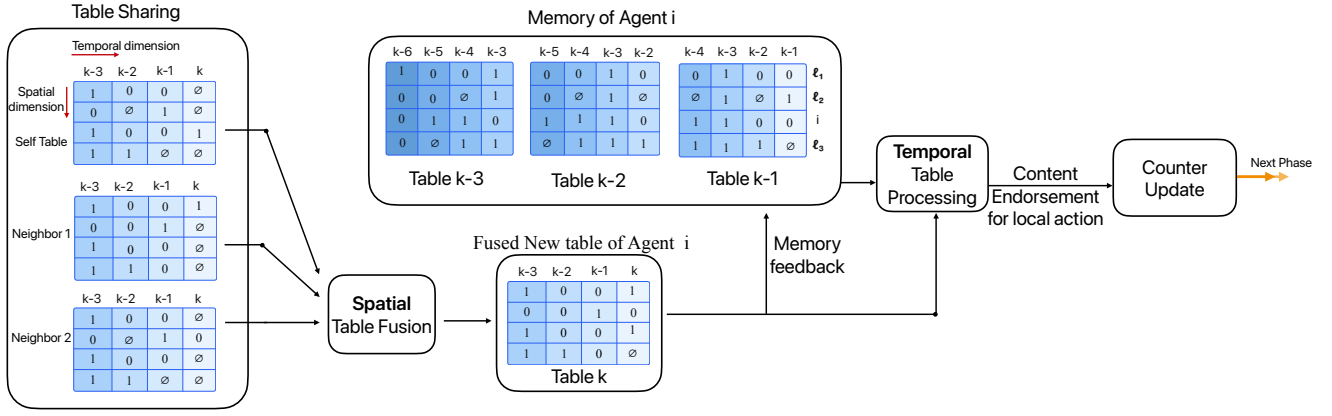


Fig. 2: A figurative illustration of spatial table fusion and temporal table processing procedures for agent $i = 1$ at phase k over the four-agent communication network depicted at the top-right corner. The agent constructs the table $o_{(k)}^i$ based on the current content/discontent signal $s_{(k)}^i$ and the previous beliefs $\{\hat{s}_{(k),(k-l)}^{i,j}\}$ about the content/discontent signals of all agents $j \in \mathcal{N}$ and at previous phases $l = k - m + 1, \dots, k - 1$. The agent receives the neighbor tables $\{o_{(k)}^j\}_{j \in \mathcal{N}_{(k)}^i}$ and fuses them with the local table $o_{(k)}^i$ to compute $\bar{o}_{(k)}^i$. Then, the agent uses $\bar{o}_{(k)}^i$ and $\{\bar{o}_{(k-l)}^i\}_{l=1}^{n-1}$ to determine whether all agents were content at phase $k - n + 1$ through majority polling.

matrix

$$o_{(k)}^i = \begin{bmatrix} \hat{s}_{(k),(k-m+1)}^{i,1} & \cdots & \hat{s}_{(k),(k)}^{i,1} \\ \vdots & \ddots & \vdots \\ \hat{s}_{(k),(k-m+1)}^{i,n} & \cdots & \hat{s}_{(k),(k)}^{i,n} \end{bmatrix} \in \{0, 1, \emptyset\}^{n \times m}. \quad (11)$$

This only creates an $n \times m$ -bit memory burden per agent, which can be further reduced, thereby yielding a scalable memory footprint. If the underlying communication networks consistently have depth more than m , then some agents may not access to the signals of all others. On the other hand, a connected network can have the depth of at most $n - 1$. Correspondingly, $m \geq n - 1$ can ensure that all agents can access to the signals of all others as long as the networks remain connected.

Spatial Table Fusion: Agents fuse the tables received by the neighboring agents to construct

$$\bar{o}_{(k)}^i = [\bar{s}_{(k),(k-l)}^i]_{l=k-m+1}^k \quad (12a)$$

$$\bar{s}_{(k),(l)}^i = [\bar{s}_{(k),(l)}^{i,j}]_{j \in \mathcal{N}}, \quad (12b)$$

as depicted in Fig. 2. Formally, we can represent the communication network at phase k by an undirected graph $G_k = (\mathcal{N}, E_{(k)})$, where $\mathcal{N}$ is the set of vertices/nodes corresponding to agents and $E_{(k)} \subset \mathcal{N} \times \mathcal{N}$ is the set of edges depicting the communication links. For example, the pair $(i, j), (j, i) \in E_{(k)}$ if agents i and j can exchange messages at phase k . We denote the neighborhood of agent i at phase k by $\mathcal{N}_{(k)}^i := \{j \in \mathcal{N} : (i, j) \in E_{(k)}\}$. Then, spatial table fusion can accept signals consistent across tables

$\{o_{(k)}^r\}_{r \in \mathcal{N}_{(k)}^i}$ and

$$\bar{s}_{(k),(l)}^{i,j} = \begin{cases} 0 & \text{if } \hat{s}_{(k),(l)}^{r,j} = 0 \forall r \in \mathcal{N}_{(k)}^i : \hat{s}_{(k),(l)}^{r,j} \neq \emptyset \\ 1 & \text{if } \hat{s}_{(k),(l)}^{r,j} = 1 \forall r \in \mathcal{N}_{(k)}^i : \hat{s}_{(k),(l)}^{r,j} \neq \emptyset, \\ \emptyset & \text{otherwise.} \end{cases} \quad (13)$$

Furthermore, at the next phase, the beliefs get updated as

$$\hat{s}_{(k+1),(k-l)}^{i,j} = \begin{cases} \bar{s}_{(k),(k-l+1)}^{i,j} & \text{if } l > 0 \\ \hat{s}_{(k+1)}^i & \text{if } l = 0 \text{ and } j = i \\ \emptyset & \text{otherwise.} \end{cases} \quad (14)$$

to construct the table $o_{(k+1)}^i$ to be shared with the neighboring agents.

Temporal Processing of Tables: Tables $\{o_{(k)}^r\}_{r \in \mathcal{N}}$ received at phase k include beliefs $\hat{s}_{(k),(k-l)}^r$ about older phases $k - l$. By considering older versions of these beliefs, agents can mitigate the impact of missing links. To this end, agents can store tables fused in memory and memory size is proportional to the table depth m .

For example, consider the example depicted in Fig. 2 with four agents. At phase k , table $\bar{o}_{(k)}^i$ includes the earliest beliefs $\hat{s}_{(k),(k-3)}^i$ while any tables $\bar{o}_{(k-l)}^i$ for $l > 3$ only consist of beliefs earlier than that, making these tables redundant to store. Hence, agents can have finite memory in which they store tables $\{\bar{o}_{(k-l)}^i\}_{l=1}^{m-1}$.

Given the tables $\{\bar{o}_{(k-l)}^i\}_{l=1}^{m-1}$ stored in the memory and the new table $\bar{o}_{(k)}^i$, agent i can use *majority polling* over $\{\bar{s}_{(k-l),(k-m+1)}^i\}_{l=0}^{m-1}$ to determine if all were content at phase $k - m + 1$. Define $b_{(l),(k)}^i = \emptyset$ if $\{j \in \mathcal{N} : \bar{s}_{(l),(k)}^{i,j} \neq \emptyset\}$ is an empty set, and otherwise

$$b_{(l),(k)}^i := \prod_{j \in \mathcal{N} : \bar{s}_{(l),(k)}^{i,j} \neq \emptyset} \bar{s}_{(l),(k)}^{i,j}, \quad (15)$$

Algorithm 1 Networked-DOEL Dynamics

Require: $\xi \in (0, 1)$, $\epsilon^i \geq 0$, $\varepsilon^i \in (0, 1)$, $C^i \geq \max_u g^i(u)$

Initialize: counter $c^i(a^i) = 0$ for all $a^i \in A^i$

Exploration	for each phase $k = 1, \dots, K$ do explore local action $a_{(k)}^i \sim \text{Uniform}(A^i)$ independently for each stage $t = (k-1)\kappa + 1, \dots, k\kappa$ do play $\tilde{a}_t^i \sim \mathcal{E}^i(a_{(k)}^i)$ $\triangleright$ Simultaneous play receive payoff u_t^i end for compute $\hat{u}_{(k)}^i(a^i)$ for all a^i as in (7) compute $\hat{u}_{*,(k)}^i = \max_{a^i} \hat{u}_{(k)}^i(a^i)$ generate content/discontent signal: $s_{(k)}^i = \begin{cases} \mathbb{I}_{\{\hat{u}_{(k)}^i(a_{(k)}^i) \geq \hat{u}_{*,(k)}^i - \epsilon^i\}} & \text{w.p. } \xi^{w^i(C^i - g^i(\hat{u}_{(k)}^i(a_{(k)}^i)))} \\ 0 & \text{o.w.} \end{cases}$ construct the table $o_{(k)}^i$ according to (11) and (14) share $o_{(k)}^i$ and receive $\{o_{(k)}^j\}_{j \in \mathcal{N}_{(k)}^i}$ construct $\bar{o}_{(k)}^i$ by fusing $o_{(k)}^i$ and $\{o_{(k)}^j\}_{j \in \mathcal{N}_{(k)}^i}$ as in (13) if $k \geq m$ and (16) holds then increment the counter $c^i(a_{(k-m+1)}^i)$ by one end if end for
Exploitation	identify $\hat{a}_{(K)}^i \in \arg \max_{a^i \in A^i} \{c_{(K)}^i(a^i)\}$ for each stage $t > \kappa K$ do play $a_t^i = \hat{a}_{(K)}^i$ $\triangleright$ Simultaneous play receive payoff u_t^i end for

which is one if all information signals $\bar{s}_{(l),(k)}^{i,j}$ are ones. Then, the agent detects network-wide content endorsement for the local explored action $a_{(k-m+1)}^i$ if

$$\sum_{l=0}^{m-1} b_{(k-l),(k-m+1)}^i \geq \frac{1}{2} \sum_{l=0}^{m-1} \mathbb{I}_{\{b_{(k-l),(k-m+1)}^i \neq \emptyset\}}. \quad (16)$$

In other words, the local action gets network-wise content-endorsement if network-wide content endorsement get detected at least half of the network-wide signals available.

Conservative table fusion accepting only signals consistent across neighboring agents and majority polling over network-wide available signals reduce the sensitivity of the dynamics to erroneous signal transmissions beyond link failures.

Algorithm 1 tabulates the Networked-DOEL dynamics for the typical agent i .

IV. ANALYTICAL RESULTS

In this section, we characterize the expected regret (5) for the Networked-DOEL dynamics over dynamic communication networks. To this end, we consider that the underlying communication network G_k at phase k gets generated *exogenously* according to the Erdős-Rényi $G(n, p)$ model. There can be a direct communication link between any pair of agents with probability $p \in (0, 1]$ independent of other pairs, previous networks, and the Networked-DOEL dynamics.

Let $q_m^i \in [0, 1]$ denote the probability that agent i can access to the signals of all others within m consecutive phases. Across m consecutive phases, there may not be any

one-length, i.e., direct, paths between agent i and some other agent j with probability $(1-p)^m$. This yields that

$$q_m^i \geq (1 - (1-p)^m)^{n-1} =: q_m \quad (17)$$

and $q_m > 0$ since $p \neq 0$. Hence, agents can access to network-wide signals $\{s_{(k)}^i\}_{i \in \mathcal{N}}$ associated with any phase k with at most m -phase delay with some *positive* probability. Indeed, information can also reach in multiple hops. However, (17) provides a tractable lower bound, growing with the table width m while decaying with the number of agents and edge probability p .

The Networked-DOEL dynamics involve in-phase exploration so that agents can identify whether the explored action is tolerable or not. However, since the other agents also conduct in-phase exploration, agents effectively identify whether the explored action is tolerable against exploration-perturbed actions of others. Therefore, we introduce the *effective payoff* $\bar{u}^i(a^i, a^{-i}) := u^i(a^i, \mathcal{E}^{-i}(a^{-i}))$ for all i and (a^i, a^{-i}) , where $\mathcal{E}^{-i}(a^{-i}) = (\mathcal{E}^j(a^j))_{j \neq i}$. Given the effective payoffs, we define the effective welfare and the effective ϵ -equilibrium set of the corresponding *effective game* $\bar{\mathcal{G}} = \langle A^i, \bar{u}^i \rangle_{i \in \mathcal{N}}$, resp., by

$$\bar{\text{SW}}(a) = \sum_{i=1}^n w^i \cdot g^i(\bar{u}^i(a)), \quad (18)$$

$$\bar{A}_\epsilon^{\text{eq}} = \left\{ a \in A : \bar{u}^i(a) \geq \max_{\bar{a}^i \in A^i} \bar{u}^i(\bar{a}^i, a^{-i}) - \epsilon^i, \forall i \right\}. \quad (19)$$

We make the following assumptions:

Assumption 1.

- (i) *The effective ϵ -equilibrium set $\bar{A}_\epsilon^{\text{eq}}$ is non-empty.*
- (ii) *There exists unique effective equilibrium $\bar{a}^* \in \bar{A}_\epsilon^{\text{eq}}$ maximizing the effective welfare $\bar{\text{SW}}$, i.e.,*

$$\bar{\text{SW}}(\bar{a}^*) > \bar{\text{SW}}(\bar{a}) \quad \forall \bar{a} \in \bar{A}_\epsilon^{\text{eq}} \text{ and } \bar{a} \neq \bar{a}^*. \quad (20)$$

- (iii) *Let $\Xi_o > 0$ denote the gap between the best and second best(s) as*

$$\Xi_o := \bar{\text{SW}}(\bar{a}^*) - \max_{\bar{a} \in \bar{A}_\epsilon^{\text{eq}} \setminus \bar{a}^*} \{\bar{\text{SW}}(\bar{a})\}. \quad (21)$$

Then, for some $\delta > 0$, we have

$$0 < \xi < (|\bar{A}_\epsilon^{\text{eq}}| + \delta)^{-1/\Xi_o}. \quad (22)$$

This assumption ensures that agents can learn the unique optimal equilibrium in a decentralized way through randomized content/discontent signals.

Assumption 2. The phase length κ is sufficiently long such that in-phase estimation errors are negligible, i.e., $\hat{u}_{(k)}^i(a_{(k)}^i) = \bar{u}^i(a_{(k)}^i, \bar{a}_{(k)}^{-i})$ and $\hat{u}_{,(k)}^i = \max_{a^i \in A^i} \bar{u}^i(a^i, \bar{a}_{(k)}^{-i})$.*

The estimates $\hat{u}_{(k)}^i$ in (7) use only agent i 's own in-phase payoffs, so they do not depend on G_k . Ordinary concentration therefore applies directly, as analyzed in [13]: Assumption 2 holds whenever κ is large enough for empirical payoffs to concentrate. We use this exact-equality form to isolate communication error in Theorem 1. In real systems, phases must be long enough relative to payoff variability;

otherwise content bits can be wrong and the bound does not apply.

The following theorem shows that the Networked-DOEL dynamics have logarithmic regret with an in-phase exploration perturbation.

Theorem 1. Consider the repeated play of a normal-form game $\mathcal{G} = \langle A^i, u^i \rangle_{i \in \mathcal{N}}$. Suppose that each agent follows Algorithm 1 and Assumptions 1 and 2 hold. Furthermore, at each phase, let the communication network be generated according to the Erdős-Rényi $G(n, p)$ model. If the exploration length be $K \in O(\log T / q_m^2)$, then the expected regret is bounded by $E[R_T] \in O(\varepsilon T + \log T)$.

The Networked-DOEL dynamics use in-phase ε -exploration to identify whether the explored actions are tolerable. However, this causes an offset on the regret. Furthermore, the exploration length $K \in O(\log T / q_m^2)$ is reversely proportional to the square of the probability lower bound q_m . As described in (17), $q_m \in (0, 1)$ increases with larger table size m , larger connectivity probability p and smaller number of agents.

Proof: We can write the expected regret as

$$E[R_T] = TSW^* - \left(\kappa \sum_{k=1}^K SW(\bar{a}_{(k)}) + (T - \kappa K)SW(\hat{a}_{(K)}) \right),$$

where $\bar{a}_{(k)} = (\mathcal{E}^i(a_{(k)}^i))_{i \in \mathcal{N}}$ is the exploration-perturbed action profile. Define $\Xi := SW^* - \min_a SW(a)$ and $\Xi_\varepsilon := SW^* - SW(\bar{a}^*)$. Then, we have

$$E[R_T] \leq \kappa K \Xi + T \Xi_\varepsilon + T \Xi \Pr(\hat{a}_{(K)} \neq \bar{a}^*). \quad (23)$$

Since the repeated play of normal-form games is a special case of stochastic games, [13, Lemma 1] yields that $0 < \Xi_\varepsilon \leq \bar{C}\varepsilon$ for certain $\bar{C} > 0$.

Next, we bound the probability $\Pr(\hat{a}_{(K)} \neq \bar{a}^*)$ from above. The challenge is that agents may not access to the network-wide signals. Furthermore, whether agent i can access to network-wide signal at phase k is not independent of having access to network-wide signals at phases $\{k - m + 1, \dots, k + m - 1\}$ since the shared tables consist of signals within a window of m phases. To address this issue, we fix agent i and consider m separate empirical averages:

$$\hat{\theta}_{(\ell),l}^{(i)}(a) := \frac{1}{\ell} \sum_{k=1}^{\ell} \mathbb{I}_{\{a_{mk+l}=a\}} \cdot s_{(mk+l)} \cdot r_{(mk+l)}^{(i)} \quad (24)$$

for $l = 0, 1, \dots, m - 1$, where $s_{(mk+l)} = \prod_j s_{(mk+l)}^j \in \{0, 1\}$ is 1 if there is network-wide content endorsement and 0 otherwise, and $r_{(mk+l)}^{(i)} \sim \text{Bernoulli}(q_m^i)$ is the indicator of whether agent i can get access to the network-wide signal possibly with some delay over the randomly selected communication networks according to the Erdős-Rényi $G(n, p)$ model. Due to the in-phase separation, $r_{(mk+l)}^{(i)}$ is independent of other $r_{(mh+l)}^{(i)}$ for $h \neq k$. Additionally having access to network-wide signal is independent of the actions taken and the signals generated.

Based on (24) and the randomness on $\mathbb{I}_{\{a_{mk+l}=a\}} \cdot s_{(mk+l)} \cdot r_{(mk+l)}^{(i)} \in \{0, 1\}$ independent across k , the Hoeffding bound yields that

$$\Pr(|\hat{\theta}_{(\ell),l}^{(i)}(a) - \theta(a)| \geq \zeta) \leq 2e^{-2\ell\zeta^2} \quad (25)$$

for $l = 0, 1, \dots, m - 1$, where

$$\theta(a) := \begin{cases} \frac{q_m^i}{|A|} \xi^{C - \overline{SW}(a)} & \text{if } a \in \bar{A}_\varepsilon^{\text{eq}} \\ 0 & \text{if } a \notin \bar{A}_\varepsilon^{\text{eq}} \end{cases}. \quad (26)$$

Observe that

$$\arg \max_a \{\theta(a)\} = \arg \max_{a \in \bar{A}_\varepsilon^{\text{eq}}} \{\overline{SW}(a)\} \quad (27)$$

since $\xi \in (0, 1)$ and $C \geq \overline{SW}(a)$ for all a . The following lemma characterizes the separation between $\theta(\bar{a}^*)$ and others $\theta(a)$ for $a \neq \bar{a}^*$.

Lemma 1. Under Assumption 1, we have

$$\theta(\bar{a}^*) - \zeta > \sum_{a \in \bar{A}_\varepsilon^{\text{eq}} \setminus \bar{a}^*} (\theta(a) + \zeta), \quad (28)$$

where

$$\zeta := \frac{q_m^i \delta}{|A| |\bar{A}_\varepsilon^{\text{eq}}|} \xi^{C - \max_{a \in \bar{A}_\varepsilon^{\text{eq}} \setminus \bar{a}^*} \{\overline{SW}(a)\}}. \quad (29)$$

Proof: The proof follows from [12, Lemma 1] with additional scaling factor q_m^i . ■

The empirical average of joint action a when there is network-wide content endorsement and agent i can access to that network-wide signal is given by

$$\hat{\theta}_{(\ell\kappa)}^{(i)}(a) := \frac{1}{m} \sum_{l=0}^{m-1} \hat{\theta}_{(\ell),l}^{(i)}(a). \quad (30)$$

Correspondingly, the local empirical average is given by

$$\hat{\theta}_{(\ell\kappa)}^i(a^i) = \sum_{\bar{a} \in A: \bar{a}^i = a^i} \hat{\theta}_{(\ell\kappa)}^{(i)}(\bar{a}). \quad (31)$$

The following proposition shows that agent i can locally identify the best action.

Proposition 1. Under Assumptions 1 and 2, if $|\hat{\theta}_{(k)}^{(i)}(a) - \theta(a)| < \zeta$, where ζ is as described in Lemma 1, then

$$\hat{\theta}_{(\ell\kappa)}^{(i)}(\bar{a}^{i,*}) > \hat{\theta}_{(\ell\kappa)}^i(a^i) \quad \forall a^i \neq \bar{a}^{i,*}. \quad (32)$$

Proof: The proof follows from [12, Proposition 1]. ■

Lastly, by (25), the triangle inequality and the union bound yield that

$$\Pr(|\hat{\theta}_{(K)}^{(i)}(a) - \theta(a)| \geq \zeta) \leq 2me^{-2K\zeta^2/\kappa}. \quad (33)$$

Correspondingly, we have

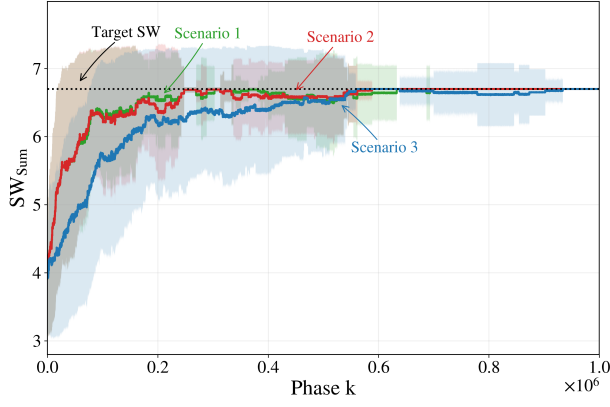
$$\Pr(|\hat{\theta}_{(K)}^{(i)}(a) - \theta(a)| \leq \zeta, \forall a \in A) \geq 1 - \beta \quad (34)$$

where $\beta := 2m|\bar{A}_\varepsilon^{\text{eq}}|e^{-2K\zeta^2/\kappa}$.

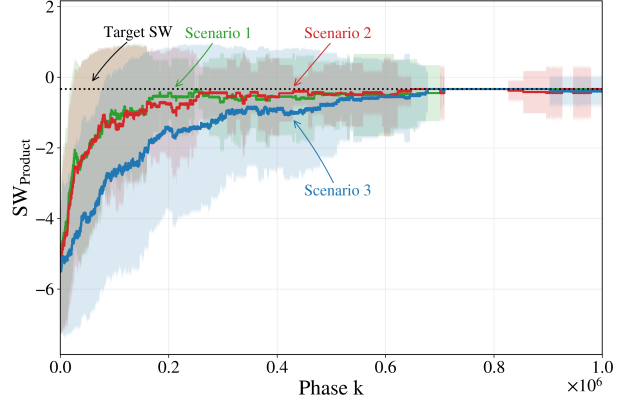
By Proposition 1 and (34), we have

$$\Pr(\hat{a}_{(K)} \neq \bar{a}^*) \leq \beta. \quad (35)$$

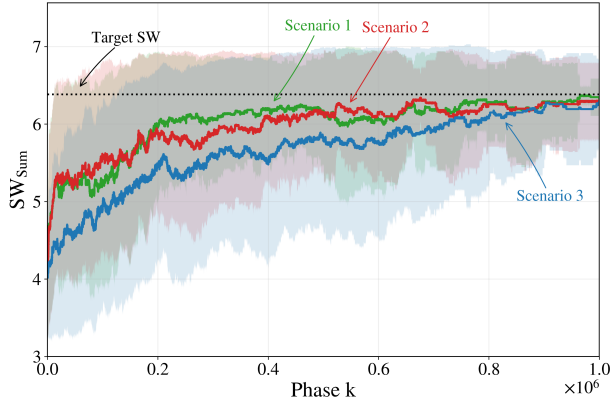
Then, choosing $K = \frac{1}{2\zeta^2} \log(4m|\bar{A}_\varepsilon^{\text{eq}}|T\zeta^2)$ yields that $\beta = \frac{1}{2T\zeta^2}$, which completes the proof. ■



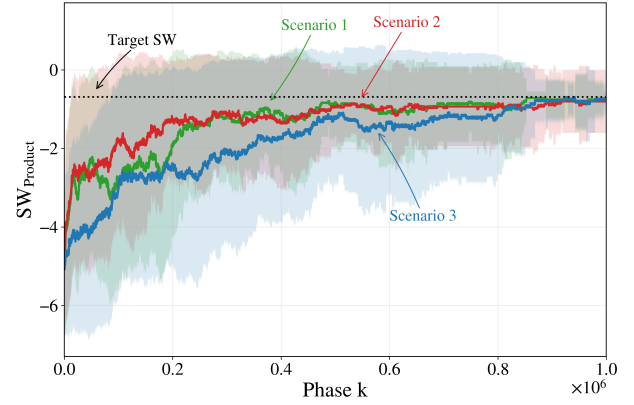
(a) Sum of Local Payoffs as Social Welfare



(b) Product of Local Payoffs as Social Welfare



(c) Sum of Local Payoffs as Social Welfare



(d) Product of Local Payoffs as Social Welfare

Fig. 3: Comparison of the proposed networked learning scheme for optimal equilibrium selection and social welfare maximization under three communication scenarios. Panels (a) and (c) use utilitarian welfare (sum of local payoffs), while panels (b) and (d) use proportional-fair welfare (sum of log-payoffs). **Scenario 1** is a fully connected baseline (ideal communication), i.e., the same communication setting as the original DOEL dynamics. **Scenario 2** and **Scenario 3** use the *restricted Erdős-Rényi* graph (edge probability $p = 0.2$, conditioned on connectivity) with per-phase independent link drops with probability 0.1 for each edge, and differ in the table depth: $m = 5$ (**Scenario 2**) versus $m = 3$ (**Scenario 3**), so that m is above versus possibly below the backbone diameter $D \in \{3, 4\}$. The dashed horizontal line indicates the target welfare level; curves show averages over 50 runs with shaded ± 1 standard deviation.

V. NUMERICAL RESULTS

We consider the repeated play of a normal-form game with $n = 7$ agents and $|A^i| = 3$ actions per agent. The payoffs are generated randomly over $[0, 1]$ once and fixed throughout all experiments. For the realized game instance, the induced equilibrium landscape contains three distinct pure Nash equilibria, each achieving a different value of the social welfare objective. In addition, the (unconstrained) social maximizer, $a^{\text{SW}} \in \arg \max_{a \in A} \text{SW}(a)$, is unique and does not coincide with any of the equilibrium action profiles, highlighting a nontrivial gap between stability (equilibrium) and efficiency (social welfare).

On this same payoff matrix we evaluate both equilibrium selection and unconstrained welfare maximization, under two welfare specifications: (i) the utilitarian case

$\text{SW}_{\text{sum}}(a) = \sum_{i=1}^7 u^i(a)$ and (ii) the proportional-fair case $\text{SW}_{\text{product}}(a) = \sum_{i=1}^7 \log(u^i(a))$ (with a small positive floor applied to avoid $\log(0)$). The same networked learning dynamics is then run in the identical environment under both cases, and performance is reported with respect to (a) the optimal equilibrium under the chosen $\text{SW}(\cdot)$ among the three equilibria, and (b) the unique social maximizer a^{SW} as a separate efficiency benchmark.

We remark that the joint action space has the size of $|A| = 2187$, i.e., there are 2187 possible joint action profiles in the stage game.

We consider a two-stage communication model which we called *restricted Erdős-Rényi*. First, a fixed Erdős-Rényi backbone graph is generated at initialization of each run and retained throughout that run, so the 50 independent runs

use different backbones. Pairs that do not share a backbone edge are never connected. The backbone is sampled so that its diameter is $D \in \{3, 4\}$. Then, at each phase, each backbone edge is independently dropped with a probability, yielding the active communication graph for that phase. Communication remains neighbor-only; however, links are unreliable over time: at the end of each phase, each edge in the base graph is independently active with a probability.

In all simulations, we use *restricted Erdős-Rényi* with edge probability $p = 0.2$ (conditioned on connectivity), and keep this base topology fixed throughout learning with each edge drop probability 0.1. Since communication occurs only once per phase (at the phase boundary), the active neighbor set is re-sampled once per phase and remains fixed during the κ within-phase interaction rounds. Hence, the set of available communication links varies across phases due to random link drops, even though the underlying base graph is time-invariant. We run the process for $K = 10^6$ phases.

To study equilibrium selection, we set the phase length to $\kappa = 250$ (so the total number of stage-game rounds is $T = \kappa K = 2.5 \times 10^8$), use $\varepsilon = 0.1$ exploration (implemented as ε -greedy), and set the near-best-response tolerance to $\epsilon^i = \epsilon = 10^{-5}$ for all agents. The semantic randomization parameter is as $\xi = 0.35$ and $\xi = 0.4$ for payoff-sum and payoff-product welfares, respectively.

To obtain the social-welfare maximizer benchmark, we do not divide the time into phases, i.e., $\kappa = 1$, $\varepsilon = 0$, and $\epsilon^i = +\infty$ (so every action is tolerable), and choose ξ small so that the induced stationary behavior concentrates on the welfare maximizer. Concretely, we use $\xi = 10^{-4}$ for payoff-sum maximization and $\xi = 1.6 \times 10^{-4}$ for payoff-product maximization. As above, we keep the same horizon $K = 10^6$.

Fig. 3 compares three communication scenarios. The plotted curves show the average achieved welfare over 50 independent runs, and the shaded regions indicate ± 1 standard deviation. The dashed horizontal line marks the target welfare level in each panel. Scenario 1 is a fully connected baseline, i.e., the same communication setting as the original DOEL dynamics [13].

A comparison with the no-communication payoff-based dynamics of [4], [5] is not included, because the instance is randomly sampled to represent a general normal-form game, and such a draw need not satisfy the structural assumptions those methods require. Scenarios 2 and 3 operate over the restricted Erdős-Rényi network with link drops, and differ only in the table depth: $m = 5$ for Scenario 2 and $m = 3$ for Scenario 3.

On a backbone of diameter D , a content bit reaches every agent within m phases if $m \geq D$, unless the same agent is dropped for D consecutive phases. Independent drop probability 0.1 makes that event rare, so Scenario 2 ($m = 5 \geq D$) recovers network-wide signals with high probability, whereas Scenario 3 ($m = 3$, which can be smaller than D) can leave some messages unrecovered. This is the experimental counterpart of q_m in Theorem 1. Unlike a fully connected network, some agent pairs are never linked,

and independent dropouts can leave an agent disconnected from the rest of the network in a given phase; reconstruction still proceeds through the table window of length m .

In the welfare-maximization setting shown in Fig. 3(a) and Fig. 3(b), the achieved welfare increases rapidly in the early phases and stabilizes near the target level. Scenario 1 converges fastest and with the smallest variability. Scenario 2 stays close to Scenario 1: even though some agents are never directly linked and may be disconnected from the network in a given phase, their information is still recovered through the table window. Scenario 3 may not recover all information, since the table depth can be smaller than the backbone diameter, and is slower and more variable, but still reaches the target.

A similar trend is observed in the equilibrium-selection setting in Fig. 3(c) and Fig. 3(d), where convergence is slower and the separation between scenarios is more pronounced. This is expected since equilibrium selection restricts the feasible set to equilibria, unlike unconstrained welfare maximization. Still, both Scenario 2 and Scenario 3 approach the target equilibrium welfare, with Scenario 2 again closer to Scenario 1.

VI. CONCLUSION

We studied decentralized learning of socially optimal equilibria in repeated normal-form games under dynamic communication networks and proposed Networked-DOEL, which combines low-bandwidth semantic signaling with time-stamped, time-stacked information exchange. We established logarithmic regret guarantees with an in-phase exploration perturbation and showed that the method remains effective under intermittent connectivity, with temporal redundancy improving robustness. Overall, the results highlight that scalable and resilient equilibrium selection is achievable with minimal communication. The dynamics remain resilient to intermittent links in the sense that a dropped or delayed bit can still be reconstructed inside a table window of length m . Promising directions for future research include developing more efficient exploration mechanisms, extending the framework to stochastic games, and incorporating function approximation methods to handle large-scale action spaces.

REFERENCES

- [1] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*, ser. Systems & Control: Foundations & Applications. Birkhäuser, 2013.
- [2] X. Cao, T. Başar, S. Diggavi, Y. C. Eldar, K. B. Letaief, H. V. Poor, and J. Zhang, "Communication-efficient distributed learning: An overview," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 851–873, 2023.
- [3] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Society for Industrial and Applied Mathematics, 1998.
- [4] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," *SIAM J. Control Optim.*, vol. 52, no. 5, pp. 2753–2770, 2014.
- [5] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," *Games Econ. Behav.*, vol. 75, no. 2, pp. 882–897, 2012.
- [6] R. Yang and M. Wu, "Learning with episodic hypothesis testing in general games: A framework for equilibrium selection," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 9, no. 3, Dec. 2025.

- [7] B. Yongacoglu, G. Arslan, and S. Yüksel, "Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information," *IEEE Trans. Autom. Control*, vol. 67, no. 10, pp. 5230–5245, 2022.
- [8] A. S. Donmez, O. Unlu, and M. O. Sayin, "Logit-Q dynamics for efficient learning in stochastic teams," *SIAM J. Control Optim.*, vol. 63, no. 5, pp. 3219–3243, 2025.
- [9] F. Christianos, G. Papoudakis, and S. V. Albrecht, "Pareto actor-critic for equilibrium selection in multi-agent reinforcement learning," *Trans. Mach. Learn. Res.*, 2023.
- [10] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 601–615, 2015.
- [11] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5872–5881.
- [12] S. T. Kiremitci, A. S. Donmez, and M. O. Sayin, "Achieving Pareto optimality in games via single-bit feedback," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2026.
- [13] —, "Decentralized optimal equilibrium learning in stochastic games via single-bit feedback," 2026, arXiv:2602.12830.
- [14] J. Zhang, K. You, and T. Başar, "Distributed discrete-time optimization in multiagent networks using only sign of relative state," *IEEE Trans. Autom. Control*, vol. 64, no. 6, pp. 2352–2367, 2019.
- [15] M. O. Sayin and S. S. Kozat, "Single bit and reduced dimension diffusion strategies over distributed networks," *IEEE Signal Process. Lett.*, vol. 20, no. 10, pp. 976–979, 2013.
- [16] —, "Compressive diffusion strategies over distributed networks for reduced communication load," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5308–5323, 2014.
- [17] X. Cao and T. Başar, "Decentralized online convex optimization with compressed communications," *Automatica*, vol. 156, p. 111186, 2023.
- [18] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.
- [19] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Distributed state estimation over time-varying graphs: Exploiting the age-of-information," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6349–6365, 2022.