Xiao Xu and Qing Zhao

# Distributed No-Regret Learning in Multiagent Systems

*Challenges and recent developments*



©ISTOCKPHOTO.COM/HAMSTER3D

Game theory is a well-established tool for studying interactions among self-interested players. Under the assumption of complete information on the game composition at each player, the focal point of game-theoretic studies has been on the Nash equilibrium (NE) in analyzing game outcomes and predicting strategic behaviors of rational players.

The difficulty in obtaining complete information in real-world applications gives rise to the formulation of repeated unknown games, where each player has access to only local information such as his own actions and utilities, but is otherwise unaware of the game composition or even the existence of opponents. In such a setting, a rational player improves his decision making through real-time interactions with the system and learns from past experiences [1]. A repeated unknown game can be viewed through the lens of distributed online learning, where the central question is whether learning dynamics of distributed players lead to a system-level equilibrium. Studies in the past few decades have revealed intriguing connections between various notions of no-regret learning at each player and certain relaxed versions of the NE at the system level [1], [2].

While one-step closer to real-world systems, repeated unknown games, in their canonical forms, often adopt idealistic assumptions in terms of the stationarity of the player population and their utilities, availability of complete and perfect feedback, full rationality of players with unbounded cognition and computation capacity, and homogeneity among players in their knowledge of the game. Many emerging multiagent systems, however, are inherently dynamic and heterogeneous, and also are inevitably limited in terms of available information and the cognition and computation capacities of the players. We present two examples.

1) *Adversarial machine learning*: Security issues are at the forefront of machine learning and deep learning research, especially in safety-critical and risk-sensitive applications. The interaction between the defender and the attacker can be modeled as a two-player game. While the player population may be small, the game is highly complex in terms of the action space, utilities, feedback models, and the available

knowledge each player has about the other. In particular, the attacker is characterized by its knowledge (how much information it has for designing attacks) and power (how often a successful attack can be launched). Both can be dynamically changing and adaptive to the defender's strategies. A full spectrum of attacker profiles has been considered, ranging from the so-called black-box model to the white-box model (i.e., an omniscient attacker). The attack process is also dynamic, often exhibiting bursty behaviors following a successful intrusion or a system malfunction. The attacker's action space can be equally diverse, including poisoning and perturbation attacks. The former targets the training phase by injecting corrupted labels and examples for the purpose of embedding wrong decision rules into the machine learning algorithm. The latter targets the blind spots of a fully trained artificial intelligence using strategically perturbed instances that trigger incorrect outputs, even when the perturbation is indiscernible to humans. In terms of utilities, the attacker's goal may be to compromise either the system's integrity (i.e., to evade detection by causing false negatives) or the availability of the system (to flood the system with false positives). See a comprehensive taxonomy of attacks against machine learning systems in [3].

2) *Transportation systems*: Route selection in urban transportation is a typical example of a noncooperative game repeated over time. The game is characterized by a large population of players that is both dynamic and heterogeneous, with vehicles leaving and joining the system and utilities varying across players and over time. The envisioned large-scale adoption of autonomous vehicles will further diversify the traffic composition. Autonomous vehicles are significantly different from human drivers in terms of decision-making rationality, access to and use of system-level knowledge, and memory and computation power. Bounded rationality is more evident in human drivers: they are likely to select a familiar route and inclined to settle for sufficing yet suboptimal options.

Complex multiagent systems, such as those detailed in the previous examples, call for new game models, regret notions, distributed learning algorithms, and techniques for analyzing game outcomes. In this article, we present representative results on distributed no-regret learning in multiagent systems. In the next section, "Distributed Learning in Repeated Unknown Games," we begin with a brief review of background knowledge on classical repeated unknown games. Subsequently, we explore four game characteristics that challenge the classical game models: dynamicity, incomplete and imperfect feedback, bounded rationality, and heterogeneity. For each characteristic, we illuminate its implications and ramifications in game modeling, notions of regret, feasible game outcomes, and the design and analysis of distributed learning algorithms. Limited by our understanding of this expansive research field and constrained by the page limit, the coverage is inevitably incomplete. Nevertheless, we hope this article provides an informative glimpse of the current landscape of this field and stimulates future research interests.

## Distributed learning in repeated unknown games
In this section, we review key concepts in game theory and highlight classical results on distributed learning in repeated unknown games.

### Static games and equilibria
An $N$-player static game is represented by a tuple $\mathcal{G}(\mathcal{N}, \mathcal{A}, u)$, where $\mathcal{N} = \{1, \ldots, N\}$ is the set of players, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ is the Cartesian product of each player's action space $\mathcal{A}_i$, and $u = (u_1, \ldots, u_N)$ are the utility functions that capture interactions among players. Specifically, the utility function $u_i$ of player $i$ encodes his or her preference toward an action. It is a mapping from the action profile $\mathbf{a} = (a_1, \ldots, a_N)$ of all players to player $i$'s reward $u_i(\mathbf{a})$.

An NE is an action profile $\mathbf{a}^* = (a_1^*, \ldots, a_N^*)$ under which no player can increase his or her reward via a unilateral deviation. Specifically, $u_i(\mathbf{a}^*) \geq u_i(a_i', \mathbf{a}_{-i}^*)$ for all $i$ and $a_i' \neq a_i^*$, where $\mathbf{a}_{-i}^*$ denotes the action profile after excluding player $i$. Due to the focus on deterministic actions (also called pure strategies), the resulting equilibrium is a pure NE. A player may also adopt a mixed strategy, which is a probability distribution $s_i$ over the action space. Correspondingly, a mixed NE is a product distribution $\mathbf{s}^* = s_1^* \times \cdots \times s_N^*$ under which the expected utility $\mathbb{E}_{\mathbf{a}^* \sim \mathbf{s}^*}[u_i(\mathbf{a}^*)]$ for every player $i$ is no smaller than that under a unilateral deviation $s_i' \neq s_i^*$ in player $i$'s strategy. A game with a finite population and a finite action space has at least one mixed NE but may not have any pure NE [4].

NE is defined under the assumption that players adopt independent strategies (note the product form of $\mathbf{s}^*$). A more general equilibrium, correlated equilibrium (CE), allows correlation across players' strategies. We note that for the equilibrium definitions introduced here, we focus on games with a finite action space. Specifically, a CE is a joint probability distribution $\mathbf{s}$ (not necessarily in a product form) satisfying $\mathbb{E}_{\mathbf{a} \sim \mathbf{s}}[u_i(a_i, \mathbf{a}_{-i}) | a_i] \geq \mathbb{E}_{\mathbf{a} \sim \mathbf{s}}[u_i(a_i', \mathbf{a}_{-i}) | a_i]$ for all $i$, $a_i$, and $a_i'$, where the expectation is over the joint strategy $\mathbf{s}$ conditioned on that the realized action of player $i$ is $a_i$. The concept of CE can be interpreted by introducing a mediator who draws an outcome $\mathbf{a}$ from $\mathbf{s}$ and privately recommends action $a_i$ to player $i$. The equilibrium condition states that no player has the incentive to deviate from the outcome of the correlated draw from $\mathbf{s}$ after his or her part is revealed. CE can be further relaxed to coarse CE (CCE), which is a joint distribution $\mathbf{s}$ satisfying $\mathbb{E}_{\mathbf{a} \sim \mathbf{s}}[u_i(\mathbf{a})] \geq \mathbb{E}_{\mathbf{a} \sim \mathbf{s}}[u_i(a_i', \mathbf{a}_{-i})]$ for all $i$ and all $a_i' \neq a_i$. Different from CE, CCE imposes an equilibrium condition that is realization independent.

The four types of equilibria exhibit a sequential inclusion relation: for a given static game, the set of CCE contains all

> A repeated unknown game can be viewed through the lens of distributed online learning, where the central question is whether learning dynamics of distributed players lead to a system-level equilibrium.

CE, the set of CE contains all mixed NE, and the set of mixed NE contains all pure NE (which, however, may not exist). The more general set of strategy profiles (i.e., allowing correlated strategies across players) in CE and CCE may lead to higher expected utilities summed over all players. CE and CCE can also be computed via linear programming, while pure and mixed NE are difficult to compute [4]. More importantly, CE and CCE can be learned through certain learning dynamics of players when a game is played repeatedly, as we discuss next. A caveat is that the set of CCE may contain highly nonrational strategies that choose only strictly dominated actions (actions that are suboptimal responses to all action profiles of the other players). See [5] for specific examples.

### Repeated unknown games and no-regret learning

A repeated game consists of $T$ repetitions of a static game (referred to as the stage game in this context). In a general definition of a repeated game [6], the stage game is parameterized by a state that affects the utility function. The state may be fixed throughout all stages or evolve over time following a Markov transition rule (the state in the next stage depends on the state and actions in the current stage). We focus on the former when discussing classical results on repeated games.

In a repeated unknown game, after taking an action $a_i^t$ (potentially randomized according to a mixed strategy) in the $t$th stage, player $i$ accrues a utility $u_i(\mathbf{a}^t)$ and observes the entire utility vector $(u_i(a_i', \mathbf{a}_{-i}^t))_{a_i' \in \mathcal{A}_i}$ for all actions $a_i'$ in his action space (we focus on a finite action space here) against the action profile $\mathbf{a}_{-i}^t$ of the other players. The actions and utilities of the other players, however, are unknown and unobservable.

From a single player's perspective, a repeated unknown game can be viewed as an online learning problem in which the player chooses actions sequentially by learning from past experiences. A commonly adopted performance measure in online learning is regret, which is defined as the cumulative reward loss against a properly defined benchmark policy with hindsight vision or certain clairvoyant knowledge about the game. In other words, the benchmark policy defines the learning objective that an online algorithm aims to achieve over time. Different benchmark policies lead to different regret measures. Two classical regret notions are external regret and internal regret, as detailed next.

Let $\pi_i$ denote the online learning algorithm adopted by player $i$. For a fixed action sequence $\{\mathbf{a}_{-i}^t\}_{t=1}^T$ of the other players, the external regret of $\pi_i$ is defined as

$$\max_{a' \in \mathcal{A}_i} \mathbb{E}_{\pi_i}\left[\sum_{t=1}^T (u_i(a', \mathbf{a}_{-i}^t) - u_i(\mathbf{a}^t))\right], \qquad (1)$$

where $\mathbb{E}_{\pi_i}$ denotes the expectation over the random action process $\{a_i^t\}_{t=1}^T$ induced by $\pi_i$. In other words, the benchmark policy in the external regret chooses the best fixed response to the other players' actions in hindsight. The internal regret of $\pi_i$ is defined as

$$\max_{a, a' \in \mathcal{A}_i} \mathbb{E}_{\pi_i}\left[\sum_{t=1}^T \mathbb{I}\{a_i^t = a\}(u_i(a', \mathbf{a}_{-i}^t) - u_i(\mathbf{a}^t))\right], \qquad (2)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. In this definition, the benchmark policy is the best hindsight modification of $\pi_i$ by swapping a single action with another throughout all stages.

An online learning algorithm $\pi_i$ is said to achieve the no-regret condition if, against all action sequences $\{\mathbf{a}_{-i}^t\}_{t=1}^T$ of the other players, the cumulative regret has a sublinear growth rate with the time horizon $T$. In other words, $\pi_i$ offers, asymptotically as $T \to \infty$, the same average reward

> The significance of no-regret learning depends on the adopted benchmark policy against which the learning algorithm is measured.

per stage as the specific benchmark policy adopted in the corresponding regret measure. No-regret learning is also referred to as Hannan consistency due to the original work in [7] and [8].

It is clear that the significance of no-regret learning depends on the adopted benchmark policy against which the learning algorithm is measured. A benchmark policy with stronger performance leads to a stronger notion of regret. In particular, the internal regret is a stronger notion than the external regret: no-regret learning under the former implies no-regret learning under the latter, but not vice versa [9].

A number of no-regret learning algorithms exist in the literature. Representative algorithms that achieve no-external-regret learning include Multiplicative Weights (MW) (also known as the Hedge algorithm) and Follow the Perturbed Leader [1]. Both are randomized policies, as randomization is necessary to achieve no-regret learning in an adversarial setting with general reward functions [1]. In particular, under the MW algorithm, each player maintains a weight $W_a(t)$ of each action $a$ at every stage $t$ based on past rewards: $W_a(t) = e^{\epsilon \sum_{\tau=1}^t r_a(\tau)} = W_a(t-1)e^{\epsilon r_a(t)}$, where $r_a(\tau)$ is the reward received under $a$ at stage $\tau$ and $\epsilon > 0$ is the learning rate. The probability of choosing $a$ in the next stage is proportional to its weight given by $W_a(t)/(\sum_{a'} W_{a'}(t))$.

For no-internal-regret learning, a representative algorithm is Regret Matching [10]. Let $R^{a \to a'}(t) = (1/t) \times \sum_{\tau=1}^t \mathbb{I}\{a_i^\tau = a\}(u_i(a', \mathbf{a}_{-i}^\tau) - u_i(\mathbf{a}^\tau))$ denote the average gain per play by switching from action $a$ to an alternative $a'$ in the past $t$ plays. In the $(t+1)$th stage, the probability of switching from the previous action $a_t$ to an alternative $a'$ is given by $(1/\epsilon)R^{a_t \to a'}(t)$, where $\epsilon > 0$ is a normalization parameter chosen to ensure a positive probability of staying with action $a_t$. Regret Matching also offers no-external-regret learning by setting the probability of selecting an action $a$ at the $(t+1)$th stage to the normalized average gain per play from playing action $a$ throughout the past $t$ plays, i.e., $R^a(t)/(\sum_{a'} R^{a'}(t))$, where $R^a(t) = (1/t)\sum_{\tau=1}^t (u_i(a, \mathbf{a}_{-i}^\tau) - u_i(\mathbf{a}^\tau))$ [10].

### System-level performance under no-regret learning

Regret captures the learning objective of an individual player. At the system level, it is desirable to know whether the dynamic behaviors of distributed players converge to an

equilibrium in some sense and whether the self-interested regret minimization promises a certain level of optimality in terms of social welfare.

For the first question, it has been shown that if every player adopts a no-external-regret learning algorithm, the empirical distribution of the sequence of actions taken by all players converges to the set of CCE of the stage game [11]. No-regret learning under the internal regret measure guarantees convergence to the more restrictive set of CE [10]. Such convergence results are, however, in terms of the empirical frequency of the players' actions rather than the actual sequence of plays. The convergence is also only to the set of equilibria, rather than to an equilibrium in the corresponding set. In fact, by treating learning in games as a dynamical system, recent studies have shown that in the continuous-time setting, the actual plays under no-regret learning algorithms (such as Follow the Regularized Leader) may exhibit cycles rather than convergence [12]. In the discrete-time setting, it has been shown that in zero-sum games, the actual plays under the MW algorithm (starting from a nonequilibrium initial strategy) diverges from every fully mixed NE [13]. For games with special structures (e.g., potential games [14] with a finite action space and bilinear smooth games [15] with a continuum of actions), however, stronger results on the convergence of the actual plays to the more restrictive set of (mixed) NE have been established.

In addition to the convergence of learning dynamics, social welfare resulting from the self-interested learning of individual players is of great interest in many applications. In (known) static games, the loss in social welfare $W(\mathbf{s}) = \mathbb{E}_{\mathbf{a} \sim \mathbf{s}} [\sum_{i=1}^{N} u_i(\mathbf{a})]$ (i.e., the system-level utility under a strategy profile $\mathbf{s}$) due to the self-interested behaviors of players is quantified by the price of anarchy (POA). It is defined as the ratio of the optimal social welfare $\mathrm{OPT} = \max_{\mathbf{s}} W(\mathbf{s})$ among all strategies to the smallest social welfare in the set of mixed NE. For repeated unknown games, a corresponding concept, price of total anarchy (POTA), is defined as

$$\frac{\mathrm{OPT}}{\min_{\mathbf{s}^1, \ldots, \mathbf{s}^T} \frac{1}{T} \sum_{t=1}^{T} W(\mathbf{s}^t)}, \qquad (3)$$

where $\mathbf{s}^1, \ldots, \mathbf{s}^T$ is the sequence of strategy profiles in the no-regret dynamics of all players.

It has been shown that in games with special structures (e.g., valid games and congestion games), no-regret learning guarantees a POTA that converges to the POA of the stage game even though the sequence of actual plays may not converge to a (mixed) NE [16]. The convergence of the POTA to the POA of the stage game implies that no-regret learning can fully negate the impact of the unknown nature of the game on social welfare. The result was later extended in [11] to a general class of games referred to as smooth games (which includes valid games and congestion games as special cases). To achieve higher social welfare, cooperation among players is necessary. For example, if every player agrees to follow a learning algorithm designed specifically for optimizing the system-level performance, the optimal action profile will be selected a high percentage of time [17].

## Dynamicity

In a dynamic repeated game, the stage game is time varying. The dynamicity may be in any of the three elements of the game composition: the set of players, the action space, and the utility functions.

> In a dynamic repeated game, the stage game is time varying. The dynamicity may be in any of the three elements of the game composition: the set of players, the action space, and the utility functions.

### Notions of regret

Dynamic unknown games call for new notions of regret to provide meaningful performance measures for distributed online learning algorithms. Specifically, the benchmark policy of a fixed single best action used in the external regret and that of a fixed single best action modification used in the internal regret can be highly suboptimal in dynamic games. As a result, achieving no-regret learning under thus-defined regret measures can no longer serve as a stamp for good performance.

A rather immediate extension of the external regret is to consider every interval of the learning horizon and measure the cumulative loss against a single best action in hindsight that is specific to each interval. This leads to the notion of adaptive regret, under which no-regret learning requires a sublinear growth of the cumulative reward loss in every interval as the interval length tends to infinity. The adaptive regret is particularly suitable for piecewise-stationary systems in which changes can be abrupt but infrequent. Classical learning algorithms such as MW can be extended to achieve no adaptive regret [18]. The key issue in algorithm design is a mechanism to discount experiences from the distant past.

Another extension of the external regret is dynamic regret, in which the benchmark policy can be an arbitrary sequence of actions, as opposed to a fixed action throughout an interval of growing length. Achieving diminishing reward loss against all sequences of actions is, however, unattainable. Constraints on either the benchmark action sequence or the reward functions are necessary for defining a meaningful measure. On the variation of the benchmark action sequence, a commonly adopted constraint in the setting with finite actions is that the benchmark sequence is piecewise-stationary with at most $K$ changes (the thus-defined regret is also referred to as the K-shifting regret). In this case, the no-adaptive-regret condition directly implies no dynamic regret [18]. With a continuum of actions, the constraint is often imposed on the cumulative distance between every two consecutive actions in the sequence, i.e., $V_T(\{a^t\}_{t=1}^{T}) = \sum_{t=1}^{T-1} \| a^{t+1} - a^t \|$. It has been shown that if the benchmark sequence is slowly varying, i.e., $V_T = o(T)$, no dynamic regret is achievable through well-designed restart

procedures [19]. The variation constraint can also be applied to the reward functions. A typical example with a continuum of actions is the sublinear "variation budget" assumption. Specifically, the cumulative variation between the reward functions in two consecutive stage games grows sublinearly in $T$, i.e., $\sum_{t=1}^{T-1} \sup_a |u_{t+1}(a) - u_t(a)| = o(T)$. Similar constraints can be imposed on the gradient $\nabla u_t(a)$ of the utility function and with the variation measured by the $L_p$ norm. See [20] and the references therein for details and corresponding no-regret learning algorithms.

The external regret and its extensions are measured against an alternative strategy of a single player. A new notion of regret—NE regret—considers a benchmark policy that is jointly determined by the strategies of all players [21]. Consider a repeated game with time-varying utility functions $\{u_i^t\}_{t=1}^T$ for each player $i$. Let $\bar{u}_i = (1/T)\sum_{t=1}^T u_i^t$ be the average utility function and $\mathbf{s}^*$ the mixed NE of the static game defined by the average utility functions $\bar{u} = (\bar{u}_1, \ldots, \bar{u}_N)$. The NE regret of player $i$ following a policy $\pi_i$ is then given by $\mathbb{E}_\pi [\sum_{t=1}^T u_i^t(\mathbf{a}^t)] - T\mathbb{E}_{\mathbf{a}^* \sim \mathbf{s}^*}[\bar{u}_i(\mathbf{a}^*)]$, where $\mathbf{a}^t$ is the action profile selected by the policies $\pi = (\pi_1, \ldots, \pi_N)$ of all players at stage $t$. No-regret learning under the NE regret ensures that each player's average reward asymptotically matches that promised by the mixed NE under the average utility functions. A centralized learning algorithm that achieves no NE regret was developed in [21] for repeated two-player zero-sum games with arbitrarily varying utility functions. Achieving no-regret learning under the measure of NE regret in a distributed setting, however, remains open.

### System-level performance

The two key measures—convergence to equilibria and POTA—for system-level performance also need to be modified to take into account game dynamics. The time-varying sequence $\{\mathcal{G}^t\}_{t=1}^T$ of stage games defines a sequence of equilibria and a sequence $\{\text{OPT}^t\}_{t=1}^T$ of optimal social welfare. The desired relation between no-regret learning dynamics at individual players and the system-level equilibria is thus in terms of tracking rather than converging. For the definition of POTA, the optimal social welfare in the numerator in (3) needs to be replaced with the average optimal social welfare $(1/T)\sum_{t=1}^T \text{OPT}^t$.

An online learning algorithm is said to successfully track the sequence of (mixed) NE in a dynamic game if the average distance between the sequence of (mixed) action profiles resulting from the algorithm and the sequence of (mixed) NE vanishes as $T$ tends to infinity. A representative study in [19] considers a game with a continuum of actions and dynamicity manifesting only in the utility functions. Under the assumptions that the sequence of NE is slow varying and the utility functions are monotonic, it was shown that learning algorithms with sublinear dynamic regret successfully track the sequence of NE. The monotonicity of the utility functions plays a key role in the analysis: it translates the closeness between the learning dynamics and the NE in terms of the cumulative reward (as in the regret measure) to the closeness

in terms of their distance in the action space (the concern of the tracking outcome).

The performance of no-regret learning in terms of social welfare was studied in [22] for games with a dynamic population of players. Specifically, in each stage, each player may independently exit with a fixed probability and is subsequently replaced with a new player with a potentially different utility function; the population size is therefore fixed, and the player set is a stationary process over time. For structural games such as first-price auctions, bandwidth allocation, and congestion games, the relation between no-adaptive-regret learning and the average optimal social welfare was examined.

Game dynamics can be in diverse forms. There is a lack of holistic understanding on the matching between regret notions and the underlying dynamics of the game. Different forms of game dynamics demand different benchmark policies to arrive at a meaningful regret measure that lends significance to the stamp of "no-regret learning" yet at the same time is attainable. Viewed from a different angle, one may pose the fundamental question: what kinds of game dynamics are tamable through distributed online learning and make no-regret learning and approximately optimal social welfare feasible?

## Incomplete and imperfect feedback

Learning and adaptation rely on feedback. Quality of the feedback in terms of completeness and accuracy thus has significant implications in no-regret learning. We explore this issue in this section.

### Incomplete feedback

Incomplete feedback stands in contrast to full-information feedback where utilities of all actions a player could have taken are observed in each stage. Incompleteness can be spatial across the action space or temporal across decision stages. In the former case, a commonly studied model is bandit feedback, where only the utility of the chosen action is revealed. In the latter, the feedback model is referred to as lossy feedback where there are decision stages with no feedback [23]. One can easily envision a more general model compounding bandit feedback with lossy feedback. Studies on this general model are lacking in the literature.

The term *"bandit feedback"* has its roots in the classical problem of multiarmed bandit [24]. The name of the problem comes from likening an archetypical single-player online learning problem to playing a multiarmed slot machine. Each arm, when pulled, generates rewards according to an unknown stochastic model or in an adversarial fashion. Only the reward of the chosen arm is revealed after each play. Due to the incomplete feedback, the player faces the tradeoff between exploration (to gather information from less-explored arms) and exploitation (to maximize immediate reward by favoring arms with a good reward history).

In a multiplayer game setting with bandit feedback, no-regret learning from an individual player's perspective can be cast as a single-player nonstochastic/adversarial bandit model

> **There is a lack of holistic understanding on the matching between regret notions and the underlying dynamics of the game.**

where the payoff of each arm/action is adversarially chosen and aggregates the interaction with the other players in the game. The concept of external regret in the game setting corresponds to the weak regret in the adversarial bandit model [25], which adopts the best single-arm policy in hindsight as the benchmark.

The MW algorithm was modified in [25] to handle the change of the feedback model from full information to bandit. Specifically, the weight $W_a(t)$ of action $a$ at time $t$ is updated as $W_a(t) = W_a(t-1) e^{\epsilon r_a(t)/p_a(t)}$ where $p_a(t)$ is the probability of selecting action $a$ at time $t$ and $r_a(t) = 0$ if $a$ is unselected. Dividing the observed reward by the corresponding probability of the chosen action ensures the unbiasedness of the observation. Quite intuitively, the price for not observing the rewards of all actions is the degradation of the regret order in the size of the action space, i.e., from $\Theta(\sqrt{\log(|\mathcal{A}|)T})$ in the full-information setting [1] to $\Theta(\sqrt{|\mathcal{A}|T})$ in the bandit setting [26].

The multiplayer bandit problem explicitly models the existence of $N$ players competing for $M$ ($M > N$) arms [27]. Originally motivated by applications in wireless communication networks where distributed users compete for access to multiple channels, this specific game model is characterized by a special form of interaction among players: a collision occurs when multiple players select the same arm, which results in utility loss. The objective of this distributed learning problem is to minimize the system-level regret over all players against the optimal centralized (collision-free) allocation of the players to the best set of arms [27]. In addition to the exploration–exploitation tradeoff in the single-player setting, this distributed learning problem under a system-level objective also faces the tradeoff between selecting a good arm and avoiding colliding with competing players. A number of distributed learning algorithms have been developed to achieve a sublinear system-level regret with respect to $T$ [27]. Recent extensions of the multiplayer bandit problem further consider the setting in which each arm offers different payoffs across players [28].

The multiplayer bandit problem is a special game model in that the players have identical action space and their interaction is only in the form of collisions when choosing the same action. In a general game setting, the impact of incomplete feedback on no-regret learning and system-level performance is largely open. One quantitative measure of the impact is the regret order with respect to the size of the action space. As mentioned previously, bandit feedback results in an additional $\sqrt{|\mathcal{A}|}$ term in the regret order, which can be significant when the action space is large. Recent work [29], [30] has shown that local communications among neighboring players in a network setting can mitigate the negative impact of bandit feedback on the regret order in $|\mathcal{A}|$. In terms of the impact on the system-level performance, it has been shown under a game model with a continuum of actions that bandit feedback degrades the convergence rate of the learning dynamics to equilibria [31].

## Imperfect feedback

Imperfect feedback refers to the inaccuracy of the observed utilities in revealing the quality of the selected actions. Recall that mixed strategies are necessary for achieving no-regret learning in the adversarial setting. The quality of a mixed strategy is characterized by the expected utility where the expectation is taken over the randomness of the strategies of all players. Referred to as expected feedback, the feedback model assuming observations on the expected utility, however, can be unrealistic. A more commonly adopted feedback model is the realized feedback, where only the utility of the realized action profile is revealed. The realized feedback can be viewed as a noisy unbiased estimate of the expected feedback where the noise is due to the randomness of players' strategies.

The noisy feedback assumes a different source of noise: it comes from the external environment and is additive to either the observed utility vectors in the so-called semibandit feedback [14] with a finite action space or the gradient of the utility functions in the first-order feedback [32] with a continuum of actions. Under the assumptions of unbiasedness and bounded variance, the issue of the additive noise can be addressed by rather standard estimation techniques and analysis. A more challenging setting is to consider nonstochastic noise due to adversarial attacks, especially in applications such as adversarial machine learning. This problem was recently studied in the single-player setting [33]. Studies in the multiagent setting are still lacking.

**Such reasoning shortcuts may be a result of limited cognition of human minds or necessitated by the available computation time and power relative to the complexity of action optimization.**

## Bounded rationality

The concept of bounded rationality was first introduced in economics [34] to provide more realistic models than the often-adopted perfect rationality that assumes the decision making of players is the result of a full optimization of their utilities. In reality, players often take reasoning shortcuts that may lead to suboptimal decisions. Such reasoning shortcuts may be a result of limited cognition of human minds or necessitated by the available computation time and power relative to the complexity of action optimization.

Cognitive limitations include the limited ability in anticipating other decision makers' strategic responses and certain psychological factors that interfere with the valuation of options. Various models exist for capturing the limitations in the players' valuation of options. For example, a player may be myopic, focusing only on the short-term reward [34]. Even with forward-thinking, a player may settle for suboptimal actions perceived as acceptable by the player [34]. The limitation in a player's ability to anticipate other players' strategies can be modeled through a cognitive hierarchy by grouping players according to their cognitive abilities and characterizing them in an iterative fashion.

Specifically, players with the lowest level of cognitive ability are grouped as the level-0 players who make decisions randomly. Level-$k$ $(k > 0)$ players are then defined iteratively as those who assume they are playing against lower-level players and anticipate the opponents' strategies accordingly. Recent work draws an interesting connection between the cognitive hierarchy model and the Optimistic Mirror Descent (OMD) algorithm for solving the saddle-point problem with applications in generative adversarial networks [36]. The saddle-point problem can be viewed as a two-player zero-sum game with a continuum of actions. The solutions to the problem correspond to the set of NE. It has been shown that the OMD algorithm guarantees a converging system dynamic to an NE in terms of the actual plays while gradient descent (GD) may lead to cycles [36]. In the language of cognitive hierarchy, players adopting GD can be regarded as level-0 thinkers in the sense that they do not anticipate the strategies of their opponents. Players adopting OMD are level-1 thinkers since they take advantage of the fact that their opponents are taking similar gradient methods, which will not lead to abrupt gradient changes between two consecutive stages [36]. Consequently, an extra gradient update is applied in OMD to accelerate learning.

Besides cognitive limitations, players are also constrained in terms of physical resources such as memory and computation power. Acquiring, storing, and processing all relevant information for decision making may be infeasible, especially in complex systems with a large action space. For example, players may only choose from strategies with bounded complexity [37] or use only recent observations in decision making due to memory constraints [38].

While models for bounded rationality abound in economics, political science, and other related disciplines, incorporating such models into distributed online learning is still in its infancy. A holistic understanding of the implications of bounded rationality in distributed online learning is yet to be gained. An intriguing aspect of the problem is that bounded rationality may not necessarily imply degraded performance. For example, in dynamic games, bounded memory of past experiences may have little effect since no-regret learning dictates that the distant past be forgotten.

## Heterogeneity

The heterogeneity of complex multiagent systems characterizes the asymmetry across players in three aspects: the available information and knowledge about the system, available actions, and the level of adaptivity to opponents' strategies. In the example of mixed traffic in urban transportation, autonomous vehicles, while likely to have greater computation power for solving decision problems, may have to obey an additional set of regulations on available actions.

In adversarial machine learning, in addition to the asymmetry on the knowledge and power, the attacker and the defender may also have different levels of real-time adaptivity to the other player's strategy. Classical regret notions, such as the external regret that assumes fixed actions of the other players, while applicable to oblivious attackers, are no longer valid under adaptive attacks. A partial solution is to adopt a new notion of policy regret defined against an adaptive adversary who assigns reward vectors based on previous actions of the player [39]. Specifically, let $u_t(\cdot; a_{1:t-1})$ denote the player's reward function determined by the adversary at time $t$, given the sequence of actions $a_{1:t-1}$ taken by the player in the past. The policy regret with reward functions $\{u_t\}_{t=1}^T$ is defined as

$$\max_{a \in \mathcal{A}} \mathbb{E}\left[ \sum_{t=1}^T u_t(a; \{a, \ldots, a\}) - \sum_{t=1}^T u_t(a_t; a_{1:t-1}) \right], \quad (4)$$

where $u_t(\cdot; \{a, \ldots, a\})$ denotes the reward function determined by the adversary if the player took actions $\{a, \ldots, a\}$ in the past. The $m$-memory policy regret is defined by assuming that the reward function depends only on the past $m$ actions of the player.

The difference between the external regret and the policy regret may not be crucial if the adversary and the player have homogeneous objectives (e.g., mixed traffic in transportation systems). It has been shown that there exists a wide class of algorithms that can ensure no-regret learning under both regret definitions, as long as the adversary is also using such an algorithm [40]. In applications such as adversarial machine learning where the adversary may be a malicious opponent, the two notions of regret are incompatible: there exists an $m$-memory adaptive adversary that can make any action sequence of the player with sublinear regret in one notion suffer from linear regret in the other [40]. A general technique for developing no-policy-regret algorithms in the single-player setting was proposed in [39]. In terms of the system-level performance, it was shown in two-player games that no-policy-regret learning guarantees convergence of the system dynamic to a new notion of equilibrium called policy equilibrium [40]. However, the understanding of policy equilibrium is limited. In games with more than two players, even the definition of policy equilibrium is unclear.

**While models for bounded rationality abound in economics, political science, and other related disciplines, incorporating such models into distributed online learning is still in its infancy.**

## Acknowledgments

## Authors

*Xiao Xu* (xx243@cornell.edu) received his bachelor's degree in electronic engineering from Tsinghua University, Beijing, in 2015 and is currently pursuing his Ph.D. degree in electrical and computer engineering at Cornell University, Ithaca, New York. His research interests include statistical online

learning, sequential decision theory, distributed learning, and game theory.

*Qing Zhao* (qz16@cornell.edu) joined Cornell University in 2015, where she is the Joseph C. Ford Professor of Engineering. Prior to that, she was a professor in the Electrical and Computer Engineering Department at the University of California, Davis. She received a Ph.D. degree in electrical engineering from Cornell University in 2001. Her research interests include sequential decision theory, stochastic optimization, machine learning, and algorithmic theory with applications in infrastructure, communications, and social-economic networks. She is a Fellow of the IEEE, a Distinguished Lecturer of the IEEE Signal Processing Society for the term of 2020–2021, a Marie Skłodowska-Curie fellow of the European Union Research and Innovation Program, and a Jubilee Chair Professor of Chalmers University during her 2018–2019 sabbatical leave. She received the 2010 IEEE Signal Processing Magazine Best Paper Award and the 2000 Young Author Best Paper Award from the IEEE Signal Processing Society.

## References

[1] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[2] H. P. Young, *Strategic Learning and Its Limits*. New York: Oxford Univ. Press, 2004.

[3] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, no. 2, pp. 121–148, 2010. doi:10.1007/s10994-010-5188-5.

[4] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[5] Y. Viossat and A. Zapechelnyuk, "No-regret dynamics and fictitious play," *J. Econ. Theory*, vol. 148, no. 2, pp. 825–842, 2013. doi: 10.1016/j.jet.2012.07.003.

[6] R. Laraki and S. Sorin, "Advances in zero-sum dynamic games," in *Handbook of Game Theory With Economic Applications*, vol. 4, H. P. Young and S. Zamir, Eds. Amsterdam, The Netherlands: Elsevier, 2015, pp. 27–93.

[7] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games*, vol. 3, M. Drescher, A. W. Tucker, and P. Wolfe, Eds. Princeton, NJ: Princeton Univ. Press, 1957, pp. 97–139.

[8] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific J. Mathematics*, vol. 6, no. 1, pp. 1–8, 1956. doi: 10.2140/pjm.1956.6.1.

[9] G. Stoltz and G. Lugosi, "Internal regret in on-line portfolio selection," *Mach. Learn.*, vol. 59, nos. 1–2, pp. 125–159, 2005. doi: 10.1007/s10994-005-0465-4.

[10] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000. doi:10.1111/1468-0262.00153.

[11] T. Roughgarden, "Intrinsic robustness of the price of anarchy," *J. ACM*, vol. 62, no. 5, 2015, Art. no. 32. doi:10.1145/2806883.

[12] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras, "Cycles in adversarial regularized learning," in *Proc. 29th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2018, pp. 2703–2717.

[13] J. P. Bailey and G. Piliouras, "Multiplicative weights update in zero-sum games," in *Proc. ACM Conf. Economics and Computation*, 2018, pp. 321–338. doi: 10.1145/3219166.3219235.

[14] A. Heliou, J. Cohen, and P. Mertikopoulos, "Learning with bandit feedback in potential games," in *Proc. Advances Neural Information Processing Systems*, 2017, pp. 6369–6378.

[15] G. Gidel, R. A. Hemmat, M. Pezeshki, R. L Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas, "Negative momentum for improved game dynamics," in *Proc. 22nd Int. Conf. Artificial Intelligence and Statistics*, 2019, pp. 1802–1811.

[16] A. Blum, M. Hajiaghayi, K. Ligett, and A. Roth, "Regret minimization and the price of total anarchy," in *Proc. 40th Annu. ACM Symp. Theory of Computing*, 2008, pp. 373–382. doi:10.1145/1374376.1374430.

[17] 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. doi:10.1137/110850694.

[18] H. Luo and R. E. Schapire, "Achieving all with no parameters: AdaNormalHedge," in *Proc. Conf. Learning Theory*, 2015, pp. 1286–1304.

[19] B. Duvocelle, P. Mertikopoulos, M. Staudigl, and D. Vermeulen, Learning in time-varying games. 2018. [Online]. Available: arXiv:1809.03066

[20] A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro, "Online optimization in dynamic environments: Improved regret rates for strongly convex problems," in *Proc. IEEE 55th Conf. Decision and Control (CDC)*, 2016, pp. 7195–7201. doi:10.1109/CDC.2016.7799379.

[21] A. R. Cardoso, J. Abernethy, H. Wang, and H. Xu, "Competing against Nash equilibria in adversarially changing zero-sum games," in *Proc. 36th Int. Conf. Machine Learning*, vol. 97, 2019, pp. 921–930.

[22] T. Lykouris, V. Syrgkanis, and É. Tardos, "Learning and efficiency in games with dynamic population," in *Proc. 27th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2016, pp. 120–129. doi:10.1137/1.9781611974331.ch9.

[23] Z. Zhou, P. Mertikopoulos, S. Athey, N. Bambos, P. W. Glynn, and Y. Ye, "Learning in games with lossy feedback," in *Proc. Advances Neural Information Processing Systems*, 2018, pp. 5140–5150.

[24] Q. Zhao, *Multi-Armed Bandits: Theory and Applications to Online Learning in Networks*. San Rafael, CA: Morgan & Claypool, 2019.

[25] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002. doi:10.1137/S0097539701398375.

[26] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. 22nd Annu. Conf. Learning Theory*, 2009, pp. 217–226.

[27] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, 2010. doi:10.1109/TSP.2010.2062509.

[28] I. Bistritz and A. Leshem, "Distributed multi-player bandits: A game of thrones approach," in *Proc. Advances Neural Information Processing Systems*, 2018, pp. 7222–7232.

[29] N. Cesa-Bianchi, C. Gentile, and Y. Mansour, "Delay and cooperation in nonstochastic bandits," *J. Mach. Learn. Res.*, vol. 20, pp. 613–650, 2019.

[30] Y. Bar-On and Y. Mansour, "Individual regret in cooperative nonstochastic multi-armed bandits," in *Proc. Advances Neural Information Processing Systems*, 2019, pp. 3110–3120.

[31] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave *N*-person games," in *Proc. Advances Neural Information Processing Systems*, 2018, pp. 5661–5671.

[32] P. Mertikopoulos and Z. Zhou, "Learning in games with continuous action sets and unknown payoff functions," *Math. Program.*, vol. 173, nos. 1–2, pp. 465–507, 2019. doi: 10.1007/s10107-018-1254-8.

[33] K.-S. Jun, L. Li, Y. Ma, and J. Zhu, "Adversarial attacks on stochastic bandits," in *Proc. Advances Neural Information Processing Systems*, 2018, pp. 3640–3649.

[34] H. A. Simon, "A behavioral model of rational choice," *Quart. J. Econ.*, vol. 69, no. 1, pp. 99–118, 1955. doi:10.2307/1884852.

[35] X. Gabaix, D. Laibson, G. Moloche, and S. Weinberg, "Costly information acquisition: Experimental analysis of a boundedly rational model," *Amer. Econ. Rev.*, vol. 96, no. 4, pp. 1043–1068, 2006.

[36] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training GANs with optimism," in *Proc. Int. Conf. Learning Representations*, 2018.

[37] M. Scarsini and T. Tomala, "Repeated congestion games with bounded rationality," *Int. J. Game Theory*, vol. 41, no. 3, pp. 651–669, 2012. doi:10.1007/s00182-011-0309-3.

[38] L. Chen, F. Lin, P. Tang, K. Wang, R. Wang, and S. Wang, "K-memory strategies in repeated games," in *Proc. 16th Conf. Autonomous Agents and MultiAgent Systems*, 2017, pp. 1493–1498.

[39] R. Arora, O. Dekel, and A. Tewari, "Online bandit learning against an adaptive adversary: From regret to policy regret," in *Proc. 29th Int. Conf. Machine Learning*, 2012, pp. 1747–1754.

[40] R. Arora, M. Dinitz, T. V. Marinov, and M. Mohri, "Policy regret in repeated games," in *Proc. Advances Neural Information Processing Systems*, 2018, pp. 6732–6741.

**SP**