The Bandit Learning beyond Stationarity under the Game Theory

Beishen Guo

{BeiShen821@gmail.com }

International Education College, Shanghai University of Finance And Economics, Shanghai, 200433, China

Abstract. It is a universal truth that classical multi-armed bandit (MAB) theory relies on the stationary assumption of reward distributions. However, this assumption is not viable in systems with multiple adaptive interacting agents where interactions endogenously redefine rewards. This paper explores bandit learning in non-stationary systems from a game-theoretic perspective, where learning in game theory, modern bandit theory, and recent advances in multi-player bandits, cooperative learning, federated decision systems, incentive design, adversarial robustness, and teamwork modeling are leveraged to propose a unifying theory of endogenous non-stationarity in bandit learning. The paper provides a conceptual framework and synthesis of strategic regret in interacting systems, stability regimes of learning systems, and how exploration-exploitation trade-offs are transformed in interactive systems.

Keywords: Endogenous Non-stationarity; Game-theoretic perspective; Strategic Regret

1 Introduction

The multi-armed bandit problem, which is also known as a standard sequential decision model, is a situational problem where a learner repeatedly chooses actions out of a finite set to get stochastic rewards, and the principle idea of this problem is to reduce the exploration and exploitation to maximize long-term cumulative reward. In classical fixed point considerations, individual arm rewards are fixed with unknown means, with the performance being cumulative regret in that respect to an optimal fixed action [1].

This fixed definition has facilitated profound theoretical study of multi-armed bandit (MAB), which, however, does not fit the real world of learning where rewards are contingent on the behavioral plans of some adaptive agents. The endogenous evolution of a reward structure by the effect of strategic interactions between concurrent learning and acting agents is often relevant in practical settings: the UCB algorithm causes violent resource congestion in online auctioning of advertising spots because of its stationary-reward assumption, and the peer-dependent hot zone hypothesis of drivers in ride-sharing schedules leads to a draconian decrease in cumulative revenue; the traditional notion of a stationary bandit model does not reflect the

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365325

peer-dependent hot zone of a driver in ride-sharing scheduling, setting back the time-varying scheduling [1].

Conventional non-stationary approaches to bandits just deal with exogenous perturbations like random reward drift and arm failure, whereas do not apply to reward evolution due to agent interaction. The game theory offers a natural analytic tool to analyze such environments: repeated games imply that agents revise their strategies at the time of observation, which constructs dynamic feedback loops that evolve the reward configuration over time [2].

This paper reexamines non-stationary bandit learning through the lens of game theory, concentrating on endogenous reward dynamics and strategic interaction, making three fundamental contributions.First of all, it formally defines the endogenous non-stationarity in multi-agent bandits and characterizes the stability regimes of adaptive learning systems.Second, synthesize cooperative, competitive, adversarial and incentive-based bandits interest models with explicit algorithmic design and quantitative performance benefits.Third, it redefines strategic regret of multi-agent interaction and characterizes the stability regimes of adaptive learning systems.

2. Theoretical bases of strategic bandit learning

2.1 Stationary Reward Model limit

The classical bandit theory with fixed reward distributions enables the algorithms to approach optimal actions after repeated sampling, and exploration exploration trade-offs only focus on identifying the fixed optimal arm of the environment. With multiple agents, the expected reward of an agent at time t depends on the overall joint action of the other agents. The rewarding functionality is endogenously evolving as agents revise their policies and the exploration of a single agent necessary changes the reward perceptions of other agents, creating a feed-back mechanism between agents and the environment that radically invalidates classical bandit algorithms, causing performance degenerative performance to be measurably reduced:

Thompson Sampling (TS): This is a class of Bayesian algorithm that provides an efficient exploration-exploitation in 1-agent problems through posterior updating [1]. TS leads to excessive concentration in the sampling actions of multi-agent resources of a small number of high-reward arms, lowering the actual system revenue to less than 30% of the theoretical one.

UCB: Upper confidence bound algorithm (classic) is incredibly congested in the shared resource issues [1]. The exploration by its optimism cause most agents to choose at the same time high-reward arms and this directly eradicates the reward advantage of these arms.

These findings confirm the essential weakness of stationary reward models in terms of being able to represent interactive endogenous non-stationarity.

2.2 Games-based learning as a dynamic model

Recurrent strategic interaction offers some form of general model of endogenous policy adjustment [2, 3]. The paper assumes that $\pi_{i,t}$ represents the mixed-strategy of agent i at time t; and where the strategies are updated by the operator: $\pi_{i,t+1} = U_i(\pi_{i,t}, H_t)$,

a definition of agent learning rule [2]. The main weakness of MAB is bandit feedback, forcing agents to just see the actual payoffs of the arms selected, with no access to counterfactual payoffs or strategies of other agents, which violates the convergence of traditional game learning policies:
Fictitious Play: The way must be completely informed of the strategies of opponents; in bandit feedback the rate of convergence is dwelt to about half because of the lack of information.

Obviously, there are cases where both can be used together, but they are also complementary, since they do not need each other.

Update on Reinforcement Learning: prone to interference by randomness of bandit rewards, converges to non-equilibrium points in games even more frequently, i.e. 60% of tested scenarios of multi-agent interaction.

Conversely, it is shown that all three rules can deterministically reach Nash equilibrium in the full information learning process with mild conditions, which forms a core emphasis on the role of bandit feedback on strategic learning processes [2].

2.3 Strategic Remorse and Adaptive Enemies

The classic cumulative regret cannot work in multi-agent interactive processes, since the adaptive opponents will bring about the dynamic variation of the optimal action of any one agent through time [1]. It describes strategic regret by agent i to quantify the real time opponent strategy dynamics:

$$R_T(i) = \sum_{t=1}^{T} (u_i(a_i^*, \pi_{-i,t}) - u_i(a_{i,t}, \pi_{-i,t})) \quad (1)$$

Where $\pi_{-i,t}$ denotes the strategies of all agents other than i at time t, and a_i^* is optimal action of agent i, given the opponents′ real-time strategies.

The book Online learning with adaptive adversaries provides an O(T) sublinear regret in the presence of limited adversarial influence, however, there are theoretical frameworks of general online convex optimization that can be applied to analyze adaptive learning phenomena [4, 5]. The most common technique of analysis of robustness in a strategic bandit process is L2-regularization-convex optimization [4]. As compared to the traditional cumulative regret, the upper bound of the strategic regret is positively related to T and the speed with which the opponent updates their strategy γ: bounded γ gives an O(T) bound, and strong adaptive adversaries give an$O(T^{2/3})$ bound [5]. Such an attribute forms the fundamental design principle behind powerful strategic bandit learning algorithms and determines the regularization weights and exploration weights to alleviate the effect of rapid adaptation by opponents.

3. Bands and Co-ordination in a Multi-player game

Competition involving shared resources is a fairly recent concept not only in economics but typically across the whole scientific community.Competition in the shared resources is a relatively new phenomenon in the domain of economics as well as generally in the entire science.

Multi-agent bandit games describe agents making choices over a common set of actions [6, 7], where the result of making the choice is congestion due to the shared resource of high-payoff arms, together with the most prevalent strategic interaction in a multi-agent bandit. The real reward of an arm given the congestion is given as:

$$r_{i,t} = \frac{\mu_{a_t}}{k_t} \tag{2}$$

where k_t is the number of agents choosing arm at time t. Congestion is classified as linear (reward inversely proportional to competing agents) and nonlinear (reward decreases as the number of agents is over and above a threshold); the latter can be found in communication channels and edge computing resources competition [6, 7]. Reward allocations become endogenous as the agents vary to avoid collisions.

The current solution to this issue is the Nash Equilibrium-based Arm Selection (NEAS) algorithm, the idea behind which is to adaptively change the exploration probability of each arm in order to bring the agent strategy distribution to the mixed strategy Nash equilibrium of the game [6]. It consists of the following main steps: (1) Dynamics estimation of the size of the agent selection of each arm through local observation; (2) Nash equilibrium distribution of strategies on the current arm set; (3) Rebalancing exploration probabilities such that it directs agents to under-selected arms. NEAS achieves a balanced distribution of system resources, which maximizes the revenue of the system by 30-40 percent as compared to exploration by independent agents over TS/UCB [6].

3.2 Mechanisms of cooperative learning.

The stability of learning in a cooperative bandit is ensured by coordinated learning through information sharing and the division of tasks in the cooperation of agents to achieve a higher efficiency in learning. Congestion is removed by the Queuing and Collaborative Knowledge (QuACK) algorithm that substitutes independent exploration with population-level sampling responsibility distribution and system regret[8]:

$$R_{T,system} = \sum_{t=1}^{T} \sum_{i=1}^{n} \left(\mu^* - \mathbb{E}[r_{i,t}]\right) \tag{3}.$$

QuACK consists of two key steps: (1) Queuing sampling: To ensure each arm is sampled by a particular agent at a particular time, eliminating the risk of simultaneous sample of arm reward distributions, agents decide the arm sampling order using a distributed queuing simulation, and (2) Global information sharing: Sampling information is exchanged using a lightweight communication protocol, and all agents update local reward estimate using global information to jointly explore arm reward distributions. Given 10 agents and 20 arms [7], QuACK can decrease system regret by over 65% relative to independent UCB exploration, and also decrease system redundant per-sample rate by 70% down to 159, decreasing exploration efficiency of multi-agent bandits by many folds.

3.3 Federated multi-agent bandits

Federated bandit learning assigns exploration to more than one client but it restricts communication, which is a common cooperative design of distributed decision systems with prohibitively limited communication constraints [9]. It trades off between velocity of learning globally and the price of localized communication, and every client will retain local estimates of rewards and periodically share data (no raw data is uploaded to a central server, so there will be privacy of the data). Its optimization problem is:

$$\min R_T + \lambda C_T \tag{4}$$

in which C_T denotes the cost of communication (in units of volume of transmission data and communication frequency), and 0 is a regularization coefficient that balances the trade-off between regret and communication cost [9].

Hierarchical Federated Bandit (HFB) architecture is an architecture that balances this trade-off wherein clients are divided into sub-clusters: complete information exchange in a cluster, as well as only exchange of reward estimate means between clusters [9]. HFB provides a cost-efficient balance between the resource-constrained distributed systems by a factor of half the system communication cost and an increase in system regret of only 10% against a traditional centralized federated architecture [9].

4. Strategy, dynamism, and affective and competitive design

4.1 The exploration should be incentive-compatible

Principal-agents bandit games bring in the hierarchic interaction where one has a designer (principal) and the learning agents, where the former sets the environment of the rewards via incentive schemes to direct the exploration-exploitation behavior of the latter in achieving the overall goals of the principal [10]. The purpose of the principal is:

$$U_P = \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t} - c(a_t)\right] \tag{5}$$

where $c(a_t)$is the cost of the incentive scheme to the principal, depending on the action t at [9]. The incentives mechanisms have to meet two fundamental criteria: individual rationality (the expected revenue of the agents when influenced by the incentives is not worse than with independent decision-making) and incentive compatibility (the optimum behavioral point of the agents with the incentives is the same as the target behavioral point of the principal) [10].

This process is most commonly applied in the Bonus Incentive-based Exploration (BIE) algorithm [9], where the bonus coefficient of each arm is decreased or increased in real time, with the high-reward but low-potential arms receiving a higher and higher bonus incentive to explore and the other way around, causing the system to be brought to a global optimum decision making. In real life, BIE is preferable to punishment and reputation mechanisms because it is simple and it boosts the rate of exploration of potential high-payoff arms by 50-70 percent in experiments of allocating resources to markets online [10].

4.2 Adversarial influence and adversarial robustness

Adversarial manipulation implements manipulations to observed rewards through strategic perturbation, and in this setup, the adversaries corrupt observed rewards, and the corrupted rewards are formulated as:

$$r'_{i,t} = r_{i,t} + \delta_{i,t} \tag{6}$$

[11]. Group of attacks: There are two forms of adversarial attacks, namely (1) Reward poisoning attack: introducing malicious noise to measured rewards in order to manipulate the agents' estimate of reward distribution; and (2) Strategy perturbation attack: introducing false agent strategies so that real agents are driven to avoid high-reward arms.

The real-time robust bandits algorithm fuses soon-state-of-anomaly robust bandits, which consist of L2-regularization and Anomaly Detection (L2-AD), which involve (1) L2-regularization of the observed rewards to shift the focus of the anomaly values; (2) Real-time extension of the malevolent perturbation in reward estimation by statistical anomaly detection and (3) Scrubbing of the remaining sources of attack sizes and probabilities. Under limited manipulation by adversaries, L2-AD represents a strategic regret upper bound of $O(T^{2/3})$, and incurs less than an 80 percent reduction in the effects of reward poisoning attacks versus unregularized algorithms [4, 11].

4.3 Teamwork instruments of strategy uncertainty

Bandit frameworks are the creation of the beliefs of agents concerning the reliability of partners in cooperative environments, and strategic uncertainty is caused by the lack of complete information about the learning capacity, execution efficiency, as well as cooperative willingness of partners [12]. The uncertainty leads to active dynamics of changes in cooperative strategies, causing endogenous non-stationarity in the reward.

The Teamwork Bandit (BTB) algorithm, is also based on Bayesian using beliefs as it updates beliefs about partner reliability with a three-step procedure: (1) Agents begin with a prior belief about partner reliability, incorporating previous knowledge; (2) Updated beliefs based on actual rewards, the Bayesian inference is performed; (3) Cooperative strategy is changed dynamically, and more resources are given to the partner with a high reliability belief and less investment is provided to the partner with a low reliability belief. BTB has been effectively used in UAV cooperative navigation and industrial robot teams operation [12]; in UAV navigation, it can be implemented to promote the success of navigation of the team to 85 out of 60 percent, depending on the beliefs of partner communication reliability.

5. Endogenous non-stationarity stability

Formally, the endogenous non-stationarity of multi-agent bandits is defined as:

$$\mu_{i,t} = \mu_i(\pi_{-i,t}) \tag{7}$$

The key idea of endogenous non-stationarity is that the expected reward of an agent at time t is the sole function of the agent of changes in the strategies used by the opponents.

The three paramount parameters that together define system convergence and stability include the rate of adaptive strategy of agents, the efficiency of information sharing within the system, and the intensity of strategic interactions, resulting in three different stability regimes with unique, quantitative convergence laws [2, 3]. Importantly, bandit feedback slows the speed at which systems converge by 30-50 percent of full-information feedback in all regimes because of incomplete information limitations [3]:

1. Competitive instability: The rate of agent adaptation is higher than the rate of system information acquisition results in the ongoing oscillation of system strategy, the amplitude of which, and system regret grows 1-2 times with a 50 percent increment in the intensity of interaction.

2. Cooperative stabilization System information sharing efficiency > agent adaptation speed: The system uses information exchange to regularly update reward estimations to the global optimum point and converges to it; nature-inspired systems QuACK and HFB converge 2 -3 times faster than independent exploration [8, 9].

3. Via adversarial destabilization: At a level of adversarial attack strength that the system is robust to, strategy changes to convergence, then change rate rises exponentially-- a 2x stronger adversarial attack will result in 4-5x faster regret growth [11].

General convergence conditions are given by learning-in-games theory [2]; however, with bandit feedback, critical delays are irreducible, which have no opportunity to be neglected when designing a stable strategic bandit algorithm.

6. Discussion

Combined learning and game theory history shows that endogenous non-stationarity is a natural phenomenon of interaction decision-making behavior: cooperative behavior minimizes uncertainty through information exchange, adversary behavior increases instability through reward distortion, and incentive design can re-model exploration behavior through reward structure adaptation. This confirms that the exploration-exploitation trade-off is not yet well understood without considering strategic interaction, and the algorithms in this paper can offer quantitative solutions to each of the interaction types.

Although the theoretical framework has made such progress, large gaps between the theory and practice remain: first, the computational complexity: most existing strategic bandit models have a complexity of $O(T \log T)$[6, 8, 9], which implies that implementing them needs significant computing resource and sequence of iterative computations, neither of which is compatible with the <100ms latency of online markets or autonomous driving; second, the homogeneous agents assumption: the current models assume that the agents are homogeneous in terms of learning speed, objective, and information gathering capabilities [2, 3].

The gaps indicate that more practical and flexible strategy bandit learning models are required to cope with the constraints of the real world.

7. Conclusion

To achieve more than the bandit learning in the stationary environment, there is a conceptual shift required of the reward model a static model to an interaction-driven model. The non-stationarity will then no longer be exogenous to the system as soon as the reward distributions are subjected to adaptive agents. Additionally, the exploration-exploitation trade-off between both agents ceases to be an independent choice but becomes a dynamical game between the agents. Two-player resource competition, two-player coordination, design incentives, adversarial system robustness against opponents, teamwork, and modeling are the basic elements of the strategic bandit learning. All these environments are explicitly supported by algorithms whose benefits in performance are quantified (NEAS to compete, QuACK to cooperate, L2-AD to be adversarial, etc.).

The derivation of game-theoretic analysis with bandit theory forms a major foundational platform of learning the phenomenon of learning in a dynamic interactive world. It overcomes the constraints of the classical stationary bandit problem not only in the end but also generalizes game-theoretic learning theory to a limited feedback environment. It is a unifying theory that combines the principles of both the bandit theory and game-theoretic learning, and it displays the laws that govern the evolution of reward structures that govern strategic interaction between multiple agents.

According to the scheme above and the gaps in the research, the four research directions of future researches are presented as follows: heterogeneous multi-agent bandit games: it will develop a theoretical framework of agents with different learning rates, objective functions, and offer adaptive algorithms that can deal with heterogeneous and multi-objective needs; endogenous non-stationarity with changing arm sets: non-stationarity will be considered in agents with time-varying sets of arms, and a low-computational-complexity algorithm to achieve approximate optimal regret is to be developed in response to the real-time.

References

[1] Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press. pp. 1-536 (2020)

[2] Fudenberg, D., Levine, D. K.: The Theory of Learning in Games. MIT Press. pp. 1-292 (1998)

[3] Perchet, V., Rigollet, P.: Multi-agent bandit learning: A game-theoretic perspective. Journal of Machine Learning Research. 22(156), pp. 1-45 (2021)

[4] Shalev-Shwartz, S.: Online Learning and Online Convex Optimization. Foundations and Trends in Machine Learning. 4(2), pp. 107-194 (2012)

[5] Arora, S., Dekel, O., Tewari, A.: Online Bandit Learning against an Adaptive Adversary. Proceedings of ICML. pp. 149-156 (2012)

[6] Gürsoy, K.: Multi-armed bandit games. Annals of Operations Research. pp. 1-24 (2024)

[7] Chen, J., Liu, Q.: Non-stationary bandit learning with strategic agents. Proceedings of NeurIPS. pp. 8901-8913 (2022)

[8] Howson, B., Filippi, S., Pike-Burke, C.: QuACK: A Multipurpose Queuing Algorithm for Cooperative k-Armed Bandits. arXiv pre-print. pp. 1-32 (2024)

[9] Fourati, F., Alouini, M.-S., Aggarwal, V.: Federated Combinatorial Multi-Agent Multi-Armed Bandits. Proceedings of ICML (PMLR). pp. 1-20 (2024)

[10] Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, El Mahdi El Mhamdi, Eric Moulines, Michael I. Jordan, Alain Durmus.: Incentivized Learning in Principal-Agent Bandit Games. arXiv stat.ML cs.GT. pp. 1-15 (2024)
[11] Zuo, J., Zhang, Z., Wang, X., et al.: Adversarial Attacks on Cooperative Multi-agent Bandits. arXiv pre-print. pp. 1-12 (2023)
[12] Alejandra López de Aberasturi-Gómez, Carles Sierra, Jordi Sabater-Mir.: Grounded Predictions of Teamwork as a One-Shot Game: A Multiagent Multi-Armed Bandits Approach. Artificial Intelligence. Vol. 338, pp. 104230 (2025)