SCIENCE CHINA
Information Sciences

• LETTER •

March 2025, Vol. 68, Iss. 3, 139201:1–139201:2
https://doi.org/10.1007/s11432-024-4267-7

Existence and uniqueness of mean field equilibrium in continuous bandit game

Xiong WANG[1], Yuqing LI[2,3*] & Riheng JIA[4*]

[1]*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[2]*School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China*
[3]*Wuhan University Shenzhen Research Institute, Shenzhen 518057, China*
[4]*School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China*

Received 25 April 2024/Revised 23 August 2024/Accepted 31 December 2024/Published online 8 February 2025

Citation Wang X, Li Y Q, Jia R H. Existence and uniqueness of mean field equilibrium in continuous bandit game. Sci China Inf Sci, 2025, 68(3): 139201, https://doi.org/10.1007/s11432-024-4267-7

Multiarmed bandit (MAB) models are widely used for sequential decision-making in uncertain environments, such as resource allocation in computer communication systems. A critical challenge in interactive multiagent systems with bandit feedback is to explore and understand the equilibrium state to ensure stable and tractable system performance. Markov games are commonly used to track state evolution and derive learning equilibria with continuous rewards but struggle to scale effectively as the number of agents increases [1]. Mean field game theory has recently gained attention as a promising solution by approximating the interactions among large-scale agents with an average effect. However, most existing mean field bandit models assume a binary reward function [2,3], which presents two critical limitations. First, the binary reward setting is overly restrictive for practical applications, such as in resource competition games, where agents typically share resources. In these cases, rewards should range continuously within the range of $[0,1]$ instead of being limited to binary outcomes of 0 or 1. Second, the assumption that an agent's state regenerates with a certain probability of deducing equilibrium oversimplifies the dynamics of typical repeated games, which involve iterative actions without state regeneration.

In this study, we, for the first time, propose a new mean field analysis framework to explore bandit games with a continuous reward function in systems encompassing enormous agents. This approach generalizes the previous binary reward model to a more universal scenario. We focus on deriving the existence and uniqueness of mean field equilibrium (MFE), which ensures the asymptotic stability of the multiagent system. To accommodate the continuous reward, we encode the learned reward into an agent state, which is then mapped to the agent's stochastic arm playing policy and updated using observed outcomes. We show that the state evolution is upper semi-continuous, enabling the existence of MFE via the fixed point theorem. Since Markov analysis suits discrete states, we transform the stochastic continuous state evolution into a deterministic ordinary differential equation (ODE) by applying stochastic approximation theory. This allows us to characterize a contraction mapping for ODE to ensure a unique MFE for the bandit game.

System model. We study the bandit game in a multiagent system involving massive agents $\mathcal{N} = \{1,\dots,N\}$. Time is discretized into slots $\{0,1,\dots,n,\dots\}$, during which each agent locally solves an MAB problem, choosing an arm from $\mathcal{M} = \{1,\dots,M\}$ in each slot n. The reward from playing an arm is influenced by the actions of all agents due to their interactions. The state of agent i is $s_n^i = [s_n^i(1),\dots,s_n^i(M)] \in \mathbb{R}^M$, where $s_n^i(j)$ is the learned reward of arm j upon to slot n. Considering large-scale agents, we adopt the distributed Hedge stationary policy for arm playing:

$$\sigma(s_n^i, j) = (1-\eta)\frac{\mathrm{Exp}\left(\beta s_n^i(j)\right)}{\sum_{k=1}^{M}\mathrm{Exp}\left(\beta s_n^i(k)\right)} + \frac{\eta}{M}, \qquad (1)$$

where β is the smoothing parameter. We refer to a_n^i as the selected arm of agent i in time slot n. Then, population profile $f_n = [f_n(1),\dots,f_n(M)]$ indicates the proportion of agents playing various arms, where $f_n(j) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\{a_n^i=j\}}$.

To address the challenges of directly analyzing the influence of individual agents' actions in large populations, we use the mean field model to approximate their interactions. Under this model, the reward $r(f_n, j)$ that an agent receives for playing arm j depends on the population profile f_n. We consider $r(f_n, j)$ as a continuous value in the range $[0,1]$ which can be easily extended to other intervals. If agent i observes a reward $r(f_n, a_n^i)$ in slot n, its state is updated:

$$s_{n+1}^i(j) = (1-\gamma_n)s_n^i(j) + \gamma_n w_n^i(j), \qquad (2)$$

where $w_n^i(j) = r(f_n, a_n^i)$ if $a_n^i = j$, and $w_n^i(j) = s_n^i(j)$ otherwise. The stepsize γ_n is a deterministic value and satisfies $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$. A typical example is $\gamma_n = \frac{1}{n+1}$, while other settings $\gamma_n = \frac{1}{(n+1)^\alpha}, \alpha \in (\frac{1}{2}, 1]$ are also valid. Let $\boldsymbol{s}_n = [s_n^1,\dots,s_n^N]$ denote the state profile, thus $\boldsymbol{s}_n \in [0,1]^{N\times M} \subset \mathbb{R}^{N\times M}$ since $r(f_n, j) \in [0,1]$.

* Corresponding author (email: li.yuqing@whu.edu.cn, rihengjia@zjnu.edu.cn)

Potential application. Our multiagent MAB model can be employed in various communication applications, as detailed in Appendix B. For example, we consider pairwise communication involving N pairs of transmitters and receivers sending data over M links. Each link is accessed opportunistically, such as via CSMA, by multiple transmitter-receiver pairs. For link j, its resource capacity is $c(j)N$, where $c(j)$ is a constant. In slot n, we assume $N(j)$ pairs transmit data over link j. Considering that the reward is a function $g(\cdot)$ of the occupied link resource, then we have $r(N,j)=g(\frac{c(j)N}{N(j)})=g(\frac{c(j)}{N(j)/N})=g(\frac{c(j)}{f_n(j)})\triangleq r(f_n,j)$.

Existence and uniqueness of MFE. Characterizing the MFE is essential for ensuring guaranteed and predictable performance in multiagent systems. Our goal is to derive the convergence of the state profile to MFE for the bandit game. Proofs for the subsequent theorems and lemma are in Appendixes C and D. Let $\Gamma:\boldsymbol{s}_n\to\boldsymbol{s}_{n+1}$ be the compound mapping of arm playing in (1) and state update in (2). We provide the definition of MFE under mapping Γ below.

Definition 1 (MFE)**.** State $\overline{\boldsymbol{s}}$ is an MFE if $\overline{\boldsymbol{s}}=\Gamma(\overline{\boldsymbol{s}})$.

From Definition 1, MFE is in fact a fixed point under Γ.

Theorem 1 (Existence of MFE)**.** There exists an MFE $\overline{\boldsymbol{s}}$ satisfying $\overline{\boldsymbol{s}}\in\Gamma(\overline{\boldsymbol{s}})$.

Since there may exist multiple MFEs under Γ, it is still hard to track which MFE the states will converge to. Empirical results in Appendix E highlight this issue by exposing suboptimal regret performance when multiple MFEs are present. This underscores the need to derive a unique MFE for more reliable performance. A unique MFE implies that there is only one fixed point. To handle bandit feedback, we use stochastic approximation to transform the discrete-time bandit game into a continuous-time ODE.

Each agent randomly selects an arm under the stationary policy, making the state evolution a stochastic process. To obtain the unique MFE for the stochastic bandit game, we employ a deterministic ODE grounded in the stochastic approximation theory [4]. The state updating of (2) is rewritten as $s_{n+1}^i(j)=s_n^i(j)+\gamma_n(w_n^i(j)-s_n^i(j))=s_n^i(j)+\gamma_n(\mathbb{E}[w_n^i(j)]-s_n^i(j)+u_n^i(j))$, where $u_n^i(j)=w_n^i(j)-\mathbb{E}[w_n^i(j)]$. Let $\boldsymbol{w}_n=[w_n^1,\ldots,w_n^N]$, and define a filtration $\{\mathcal{F}_n\}_{n\geqslant 0}$ generated by stochastic processes $\{\boldsymbol{s}_n,\boldsymbol{w}_n\}_{n\geqslant 0}$. Clearly, $\mathcal{F}_n$ is a σ-algebra with $\mathcal{F}_n\subset\mathcal{F}_{n+1}$. As a result, the process $\{u_n^i\}_{n\geqslant 0}$ is a martingale, and $\mathbb{E}[u_n^i(j)|\mathcal{F}_n]=0,\forall j\in\mathcal{M}$. The state profile $\boldsymbol{s}_n$ is defined at the discrete timescale, while the asymptotic pseudotrajectory involves two continuous-time processes. We introduce a continuous-time interpolated process for $\boldsymbol{s}_n$ to link $\boldsymbol{s}_n$ to $\boldsymbol{s}_t$. Let $\tau_0=0$ and $\tau_n=\sum_{k=1}^{n-1}\gamma_k,\forall n\geqslant 1$. Define the interpolated process $\tilde{s}_{\tau_n+h}^i=s_n^i+h\frac{s_{n+1}^i-s_n^i}{\tau_{n+1}-\tau_n},0<h<\gamma_n,\forall i\in\mathcal{N}$. By analyzing the convergence of $\boldsymbol{s}_n$ via a deterministic process $\boldsymbol{s}_t$, the interpolated process $\tilde{\boldsymbol{s}}_t$ is an asymptotic pseudotrajectory.

Lemma 1 (Asymptotic pseudotrajectory)**.** Let $\tilde{s}_t^i$ denote the interpolated process of state $s_n^i=[s_n^i(1),\ldots,s_n^i(M)]$ for agent i. Then, $\tilde{s}_t^i$ is an asymptotic pseudotrajectory for the solution s_t^i to the ODE $\frac{\mathrm{d}s_t^i}{\mathrm{d}t}=\mathbb{E}[w_t^i|\boldsymbol{s}_t]-s_t^i$.

This lemma states that if the ODE of $\boldsymbol{s}_t$ solely converges to a single fixed point, there is a unique MFE for $\boldsymbol{s}_n$, since the asymptotic pseudotrajectory ensures that the stochastic approximation error will reduce to zero. The mapping Γ now indicates that the ODE $s_t^i(j)$ is $\frac{\mathrm{d}s_t^i(j)}{\mathrm{d}t}=\sigma(s_t^i,j)\big(r(f(\boldsymbol{s}_t),j)-s_t^i(j)\big)$, where $r(f(\boldsymbol{s}_t),j)$ is the expected reward over the state profile $\boldsymbol{s}_t$ since the population profile $f(\boldsymbol{s}_t)$ is mapped from $\boldsymbol{s}_t$ by the stochastic stationary policy. Next, we first assume that the reward function acts as a contraction mapping to obtain the unique MFE, and then derive conditions for this contraction property.

Theorem 2 (Uniqueness of MFE)**.** Suppose $r(f(\boldsymbol{s}_t),j)$ is a $||\cdot||_\infty$-contraction in state $\boldsymbol{s}_t$. Then the fixed point $\overline{\boldsymbol{s}}$ is the unique MFE for the bandit game. $\overline{\boldsymbol{s}}$ is also the global attractor for the ODE, and $\boldsymbol{s}_t$ converges to the global attractor $\overline{\boldsymbol{s}}$ with exponential rate.

We also discuss the convergence rate of the discrete-time $\boldsymbol{s}_n$ in Appendix C.4. Suppose that the reward function $r(f_n,j)$ is θ-Lipschitz continuous in the population profile f_n with regard to $||\cdot||_1$-norm: $|r(f_n,j)-r(f_n',j)|\leqslant\theta||f_n-f_n'||_1$. We now characterize the contraction mapping condition.

Theorem 3 (Contraction mapping condition)**.** If parameters β,η in (1) and θ for the Lipschitz continuity satisfy the condition $4\theta(1-\eta)\beta<1$, then the reward function $r(f(\boldsymbol{s}_t),j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$.

The contraction condition is less stringent when $r(f(\boldsymbol{s}_t),j)$ is linear in $f(\boldsymbol{s}_t)$, as explained in Appendix C.6.

State change and model extension. We now calculate the cumulative state change to analyze regret as an agent state encodes the learned reward. After an agent plays an arm, its state is updated by (2). The state change is represented as an M-length vector $\Delta s_n^i=\gamma_n(w_n^i-s_n^i)$. For more results and proofs, refer to Appendix D.

Theorem 4 (Cumulative state change)**.** For any agent i and an arbitrary arm j, we have $\beta s_0^i(j)+\sum_{n=0}^T\beta\Delta s_n^i(j)-\ln(\sum_{j=1}^M\mathrm{Exp}(\beta s_0^i(j)))\leqslant\sum_{n=0}^T[\frac{1}{1-\eta}\beta(\sigma(s_n^i)-\frac{\eta}{M}\mathbf{1})\Delta s_n^i+\frac{1}{1-\eta}(e-2)\beta^2\sigma(s_n^i)\cdot(\Delta s_n^i)^2]$, where $\mathbf{1}$ is an M-length vector with each element equal to 1.

Model extensions are presented in Appendixes D.2–D.4.

Evaluation and conclusion. Comprehensive evaluation results can be found in Appendix E. To summarize, we validate our theoretical findings by demonstrating the existence and uniqueness of MFEs and by evaluating the empirical regrets under diverse reward function types.

In essence, this work proposes a general mean field framework to analyze the multiagent bandit game with a continuous reward system [5]. We confirm the existence and uniqueness of MFE to ensure a guaranteed performance.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2022ZD0115301), Fundamental Research Funds for the Central Universities (Grant No. 2042023kf0120), National Natural Science Foundation of China (Grant Nos. 62272417, 62202185, 62302343), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515110396).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Héliou A, Mertikopoulos P, Zhou Z. Gradient-free online learning in continuous games with delayed rewards. In: Proceedings of ICML, Vienna, 2020. 4172–4181
2 Gummadi R, Johari R, Schmit S, et al. Mean field analysis of multi-armed bandit games. 2013. https://ssrn.com/abstract=2045842
3 Zhao Z, Liu A L. Intelligent demand response for electricity consumers: a multi-armed bandit game approach. In: Proceedings of IEEE ISAP, San Antonio, 2017. 1–6
4 Benaim M. Dynamics of stochastic approximation algorithms. In: Séminaire de Probabilités XXXIII. Berlin: Springer, 1999. 1–68
5 Wang X, Jia R. Mean field equilibrium in multi-armed bandit game with continuous reward. In: Proceedings of IJCAI, 2021