SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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*

Appendix A System objective

We first explain the procedure of the bandit game. Denote $\boldsymbol{s}_n = [s_n^1, \cdots, s_n^N]$ as the state profile, i.e., the states of all agents, thus $\boldsymbol{s}_n \in [0,1]^{N\times M} \subset \mathbb{R}^{N\times M}$ since $r(f_n, j) \in [0,1]$. In time slot n, each agent i follows the stationary policy of Eq. (1) to map its state s_n^i to the probability distribution $\sigma(s_n^i)$, and pulls an arm a_n^i accordingly. Then, agent i observes a realized reward $r(f_n, a_n^i)$ coupled by the population profile f_n. With these observations, the state profile $\boldsymbol{s}_n$ is updated following the rule described in Eq. (2), and the bandit game moves to the next time slot. We show the whole process in Algorithm A1.

Algorithm A1 Mean field bandit game

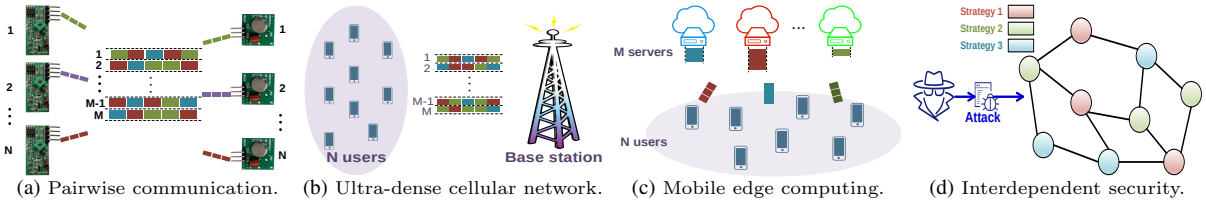
Data: Parameters η, β, stepsize γ_n
Initialize the state profile $\boldsymbol{s}_0$;
for $n = 0, 1, 2, \cdots$ **do**
  for $i \in \mathcal{N}$ **do**
    Calculate the playing policy $\sigma(s_n^i)$ according to Eq. (1);
    Choose an arm $a_n^i \sim \sigma(s_n^i)$;
    Observe the realized reward $r(f_n, a_n^i)$;
    Update the state according to Eq. (2);
  end
end

Our main objective is to analyze the convergence of states $\boldsymbol{s}_n$ to steady values, particularly to derive the *existence and uniqueness of MFE*, i.e., the multi-agent system will eventually be stable. Also, we will obtain the cumulative state change during agent playing arms, which in fact contains the regret information since each state essentially implies an agent's learned reward. However, the consideration of system stability may contrast regret minimization in that the state in traditional tight-regret policies is often the cumulative reward, rather than the learned reward in this paper. As a general Markov model is not available due to the continuous reward function and the state evolution is stochastic, we will apply a mean field analysis to transform the bandit game into an ODE, which can facilitate characterizing the convergence issue through a deterministic method. That is, our work fills the gap of analyzing interactions among numerous agents with limited feedback. The main notations are illustrated in Table A1.

Table A1 Notation

Notation	Description
N, M	# of agents, # of arms
$s_n^i, \boldsymbol{s}_n\}$	state of agent i, state profile
$f_n, r(f_n, a_n^i)$	population profile, reward function
$\sigma(s_n^i)$	arm playing policy or probability

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



(a) Pairwise communication. (b) Ultra-dense cellular network. (c) Mobile edge computing. (d) Interdependent security.

Figure B1 Potential applications.

Appendix B Potential applications in computer communication

MAB problem is widely explored in computer communication and services computing area. Aside from the pairwise communication, we next exemplify our multi-agent MAB model with its potential applications in cellular network, mobile edge computing and network security. Concretely, we demonstrate that the reward is indeed a function of the population profile as agents are regarded symmetric and asymptotically independent in the mean field model.

Cellular network. In the ultra-dense cellular network, shown in Figure B1(b), mobile users communicate with a base station through wireless channels. Interference occurs if multiple users simultaneously select the same channel. We suppose there are M channels (arms) and N users, where the user density is $\frac{\lambda}{N}$. In time slot n, the transmission power for each user is p. Besides, $N(j)$ users choose channel j with bandwidth $w(j)$, channel gain $h(j)$ and channel noise variance σ^2 [1]. The reward is a function $g(\cdot)$ of the transmission rate, which is obtained as:

$$r(N,j) = g\Big(w(j)\log_2\Big(1+\frac{ph^2(j)}{\sigma^2+\frac{\lambda}{N}N(j)ph^2(j)}\Big)\Big) = g\Big(w(j)\log_2\Big(1+\frac{ph^2(j)}{\sigma^2+\lambda f_n(j)ph^2(j)}\Big)\Big) \triangleq r(f_n,j). \tag{B1}$$

Mobile edge computing. Due to the limited local resource, mobile users will send compute-intensive tasks to nearby edge servers, who can assist them in processing the tasks. Considering a general case in Figure B1(c), there are N users and M edge servers (arms). For edge server j, its computing capacity is $c(j)N$. In time slot n, each user asks for v units of computing resource from a server, and needs to pay b for using one unit of resource. We denote the number of users choosing server j as $N(j)$. In [2], a user's profit is proportional to the allocated resource, then the reward is:

$$r(N,j) = c(j)N\frac{v}{N(j)v} - bv = \frac{c(j)}{f_n(j)} - bv \triangleq r(f_n,j). \tag{B2}$$

Network security. In an interdependent security game [3], many agents have to make individual decisions about their security actions, like deploying antivirus filters, firewalls, *etc.* We consider N agents continuously participate in the security game, and have M different security strategies (arms) to choose from, as illustrated in Figure B1(d). In time slot n, the number of agents choosing strategy j is $N(j)$. The security level of agent i depends on not only his own choices but others' actions. Specifically, the security level improves when more agents choose the same security strategy due to their interdependence. In [4], the reward is a function $g(\cdot)$ of security breach probability, which is determined by the fraction of agents playing the action, i.e., $r(N,j) = g\left(\frac{N(j)}{N}\right) \triangleq r(f_n,j)$.

Appendix C Existence and uniqueness of MFE

In this paper, we use MFE to indicate the *steady state* of a multi-agent system, which is a bit different from the concept of traditional Nash equilibrium in game theory. Nonetheless, both MFE and Nash equilibrium share the spirit of ensuring the system stability. Unlike [5], which uses mean field analysis to explore the equilibrium state among large populations with complete system information, our MFE considers the bandit feedback, introducing a distinct perspective on equilibrium analysis.

Appendix C.1 Proof of Theorem 1

The state profile $\boldsymbol{s}_n$ is in a nonempty, compact, and convex set $[0,1]^{N\times M}$. The mapping Γ maps $\boldsymbol{s}_n$ to $\boldsymbol{s}_{n+1} \in [0,1]^{N\times M}$ which is a nonempty, closed, convex subset of $[0,1]^{N\times M}$. To demonstrate there exists a fixed point, we prove that Γ is upper semi-continuous.

Let $\mathcal{P}$ be the set of the population profile f_n. For any state profile $\boldsymbol{s}_n$, we have $\Gamma_1(\boldsymbol{s}_n) \in \mathcal{P}$ based on the playing process. From Eq. (2), if agent i plays arm j, then its state corresponding to this arm after updating satisfies:

$$s^i_{n+1}(j) \in \bigcup_{f_n\in\mathcal{P}} \left\{(1-\gamma_n)s^i_n(j) + \gamma_n r(f_n,j)\right\}. \tag{C1}$$

Suppose there are arbitrary sequences $\boldsymbol{x}_n \in [0,1]^{N\times M}$ and $\boldsymbol{y}_n \in [0,1]^{N\times M}$ such that:

$$\lim_{n\to\infty}\boldsymbol{x}_n \to \overline{\boldsymbol{x}}, \quad \lim_{n\to\infty}\boldsymbol{y}_n \to \overline{\boldsymbol{y}},$$

and $\boldsymbol{y}_n \in \Gamma(\boldsymbol{x}_n)$. Given a state profile, the set $\mathcal{P}$ is determined. As $r(f_n,j)$ is continuous in f_n, we can claim that $\overline{\boldsymbol{y}} \in \Gamma(\overline{\boldsymbol{x}})$ according to Eq. (C1), i.e., Γ is upper semi-continuous. Applying the Kakutani fixed-point theorem, there exists a fixed point (MFE) $\overline{\boldsymbol{s}}$ under the mapping Γ.

Appendix C.2 Proof of Lemma 1

Appendix C.2.1 *Asymptotic pseudotrajectory*

Before we delve into the main proof, we first introduce the concept of the *asymptotic pseudotrajectory* [6]. Suppose a continuous mapping Φ on the space $\mathbb{R}^{N\times M}$ is a semiflow:

$$\Phi : \mathbb{R}_+ \times \mathbb{R}^{N\times M} \to \mathbb{R}^{N\times M},$$
$$(t,\boldsymbol{s}) \to \Phi(t,\boldsymbol{s}) := \Phi_t(\boldsymbol{s}),$$

such that

$$\Phi_0 = \text{Identity}, \quad \Phi_{t+h} = \Phi_t \circ \Phi_h.$$

In fact, $\Phi_t(\boldsymbol{s})$ can be interpreted as the state evolution process at the continuous timescale. We use $\boldsymbol{s}_t = [s_t^1, \cdots, s_t^N]$ to denote the state profile at continuous time t, which is also considered as a mapping from non-negative real numbers $t \in \mathbb{R}_+$ to the set $\mathbb{R}^{N\times M}$.

Definition 1. A continuous mapping $\boldsymbol{s} : \mathbb{R}_+ \to \mathbb{R}^{N\times M}$ is an asymptotic pseudotrajectory for Φ if:

$$\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} d\left(\boldsymbol{s}_{t+h}, \Phi_h(\boldsymbol{s}_t)\right) = 0, \forall T > 0, \tag{C2}$$

where $d(\cdot)$ is a distance measure.

From this definition, if $\boldsymbol{s}_{t+h}$ is an asymptotic pseudotrajectory for $\Phi_h(\boldsymbol{s}_t)$, then they have the *same convergence property*, with the gap diminishing to zero.

Appendix C.2.2 *Main proof*

We first demonstrate $\lim_{n\to\infty}\gamma_n = 0$. Suppose that $\gamma_n \nrightarrow 0$, and then there exists a value $\epsilon > 0$ such that $\gamma_n \geqslant \epsilon, \forall n$. Hence, $\sum_{n=0}^{K-1}\gamma_n^2 \geqslant K\epsilon^2 \to \infty$ as $K \to \infty$, which is contradictory to stepsize condition. Let $m(t) = \sup\{n \geqslant 0, t \geqslant \tau_n\}$. According to Robbins-Monro theorem [6], to prove $\bar{s}_t^i$ is the asymptotic pseudotrajectory for the above ODE in Lemma 1, we only need to show that the discrete processes s_n^i, u_n^i satisfy the following two conditions.

1) For all $T > 0$, $\lim_{n\to\infty}\sup\{||\sum_{l=n}^{k-1}\gamma_l u_l^i||_2^2 : k = n+1, \cdots, m(\tau_n + T)\} = 0$. From the Hölder's inequality, we have:

$$||\sum_{l=n}^{k-1}\gamma_l u_l^i||_2^2 \leqslant \sum_{l=n}^{k-1}\gamma_l^2||u_l^i||_2^2.$$

Since $\lim_{n\to\infty}\{\sum_{l=n}^{k-1}\gamma_l^2 : k = n+1, \cdots, m(\tau_n+T)\} = 0$, we are left to show $||u_l^i||_2^2$ is finite. In fact, $u_l^i = w_l^i - \mathbb{E}[w_l^i]$, and then $||u_l^i||_2^2 = ||w_l^i - \mathbb{E}[w_l^i]||_2^2 \leqslant ||w_l^i||_2^2 + ||\mathbb{E}[w_l^i]||_2^2 \leqslant 2M$. As a result, we claim condition 1) holds.

2) $\sup_n ||s_n^i||_2^2$ is bounded. This is naturally satisfied due to $s_n^i(j) \in [0,1], \forall j \in \mathcal{M}$.

Appendix C.3 Proof of Theorem 2

In the following, we sequentially prove the unique MFE, global attractor and convergence rate.

Unique MFE. Because the reward function $r(f(\boldsymbol{s}_t), j)$ is a contraction mapping in the state profile $\boldsymbol{s}_t$, $r(f(\boldsymbol{s}_t), j) - s_t^i(j) = 0$ has only one fixed point $\overline{\boldsymbol{s}}$. By leveraging the pseudotrajectory in Lemma 1, we obtain that $\overline{\boldsymbol{s}}$ is also the unique MFE for the bandit game.

Global attractor. The MFE $\overline{\boldsymbol{s}}$ is a global attractor when $\boldsymbol{s}_t$ converges to $\overline{\boldsymbol{s}}$ from any initial point. To show this, we construct Lyapunov function $V(\boldsymbol{s}_t) = ||\boldsymbol{s}_t - \overline{\boldsymbol{s}}||_\infty$. W.l.o.g., we assume $V(\boldsymbol{s}_t)$ gets its maxima at $s_t^i(j)$, i.e., $V(\boldsymbol{s}_t) = |s_t^i(j) - \overline{s}^i(j)|$. If $V(\boldsymbol{s}_t)$ eventually degenerates to 0 regardless of the initialized value of $\boldsymbol{s}_t$, then $\overline{\boldsymbol{s}}$ is a global attractor. Therefore, we make a classified discussion: $s_t^i(j) > \overline{s}^i(j)$ and $s_t^i(j) \leqslant \overline{s}^i(j)$. For the case $s_t^i(j) > \overline{s}^i(j)$, we have $V(\boldsymbol{s}_t) = s_t^i(j) - \overline{s}^i(j)$. Take the derivative over time t and use the fact that $r(f(\overline{\boldsymbol{s}}), j) = \overline{s}^i(j)$:

$$\begin{aligned}
\frac{dV(\boldsymbol{s}_t)}{dt} &= \frac{d\left(s_t^i(j) - \overline{s}^i(j)\right)}{dt} \\
&= \sigma(s_t^i, j)\left[r(f(\boldsymbol{s}_t), j) - s_t^i(j) - r(f(\overline{\boldsymbol{s}}), j) + \overline{s}^i(j)\right] \\
&= \sigma(s_t^i, j)\left[r(f(\boldsymbol{s}_t), j) - r(f(\overline{\boldsymbol{s}}), j) - (s_t^i(j) - \overline{s}^i(j))\right] \\
&\leqslant \sigma(s_t^i, j)\left[C_1||\boldsymbol{s}_t - \overline{\boldsymbol{s}}||_\infty - |s_t^i(j) - \overline{s}^i(j)|\right] \\
&= \sigma(s_t^i, j)(C_1 - 1)|s_t^i(j) - \overline{s}^i(j)| \\
&= \sigma(s_t^i, j)(C_1 - 1)V(\boldsymbol{s}_t),
\end{aligned} \tag{C3}$$

where $C_1 < 1$ because of the contraction mapping. As a result, $\frac{dV(\boldsymbol{s}_t)}{dt} \leqslant 0$ and $\frac{dV(\boldsymbol{s}_t)}{dt} = 0$ only at $\overline{\boldsymbol{s}}$. For the case $s_t^i(j) \leqslant \overline{s}^i(j)$, sharing the same spirit, one can obtain:

$$\frac{dV(\boldsymbol{s}_t)}{dt} = \frac{d(\overline{s}^i(j) - s_t^i(j))}{dt} \leqslant \sigma(s_t^i, j)(C_1 - 1)V(\boldsymbol{s}_t). \tag{C4}$$

Therefore, the fixed point $\overline{\boldsymbol{s}}$ is a global attractor for the ODE.

Convergence rate. Based on Eqs. (C3) and (C4), Lyapunov function satisfies the condition $\frac{dV(\boldsymbol{s}_t)}{dt} \leqslant \sigma(s_t^i, j)(C_1 - 1)V(\boldsymbol{s}_t)$, which also means that:

$$\frac{dV(\boldsymbol{s}_t)}{V(\boldsymbol{s}_t)} \leqslant \sigma(s_t^i, j)(C_1 - 1)dt.$$

Hence, we can easily derive that:

$$V(\boldsymbol{s}_t) \leqslant C_2\text{Exp}[\sigma(s_t^i, j)(C_1 - 1)t], \tag{C5}$$

where C_2 is a constant. In other words, $\boldsymbol{s}_t$ converges to $\overline{\boldsymbol{s}}$ exponentially fast.

Appendix C.4 Convergence rate of discrete-time state

Theorem 2 points out that the population profile $\boldsymbol{s}_t$ will eventually approach the unique MFE $\overline{\boldsymbol{s}}$ with an exponential rate. In line with the asymptotic pseudotrajectory in Lemma 1, we also have $\boldsymbol{s}_n \to \overline{\boldsymbol{s}}$ when $n \to \infty$. Next, we discuss the convergence rate of the discrete-time $\boldsymbol{s}_n$ using the distance $||\boldsymbol{s}_n - \overline{\boldsymbol{s}}||_\infty$. Recall from the condition for the stepsize γ_n, we can set $\gamma_n = \frac{1}{(n+1)^\alpha}, \alpha \in \left(\frac{1}{2}, 1\right]$.

Corollary 1. Suppose that the reward function $r(f(\boldsymbol{s}_n), j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_n$. Denote the distance $e_n = ||\boldsymbol{s}_n - \overline{\boldsymbol{s}}||_\infty$:

1) if $\alpha \in \left(\frac{1}{2}, 1\right)$, given $n = \Omega\left(\left(\frac{\ln\frac{1}{\delta\epsilon}}{\epsilon^2}\right)^{\frac{1}{\alpha}} + \left(\ln\frac{1}{\epsilon}\right)^{\frac{1}{1-\alpha}}\right)$, then $e_n \leqslant \epsilon$ with probability at least $1-\delta$;

2) if $\alpha = 1$, given $n = \Omega\left((2+\Psi)^{\ln\frac{1}{\epsilon}} \frac{\ln\frac{1}{\delta\Psi\epsilon}}{\Psi^2\epsilon^2}\right)$, then $e_n \leqslant \epsilon$ with probability at least $1-\delta$ for any positive constant Ψ.

Proof. Still express the state evolution as $s^i_{n+1} = (1-\gamma_n)s^i_n + \gamma_n(\mathbb{E}[w^i_n] + u^i_n)$ with $u^i_n = w^i_n - \mathbb{E}[w^i_n]$. We know that u^i_n is a martingale, and $\mathbb{E}[u^i_n(j)] = 0, |u^i_n(j)| \leqslant 1$. According to Theorems 2 and 3 in [7], we only need to prove $||\mathbb{E}[\boldsymbol{w}_n] - \overline{\boldsymbol{s}}||_\infty \leqslant C||\boldsymbol{s}_n - \overline{\boldsymbol{s}}||_\infty$ with $C \in [0,1)$, to draw conclusions 1) and 2). Assume that $||\mathbb{E}[\boldsymbol{w}_n] - \overline{\boldsymbol{s}}||_\infty = |\mathbb{E}[w^i_n(j)] - \overline{s}^i(j)|$. Considering $\mathbb{E}[w^i_n(j)] = \sigma(s^i_n, j)r(f(\boldsymbol{s}_n), j) + (1-\sigma(s^i_n, j))s^i_n(j)$, and $\overline{\boldsymbol{s}}$ is a fixed point with $r(f(\overline{\boldsymbol{s}}), j) = \overline{s}^i(j)$, we have:

$$
\begin{aligned}
&||\mathbb{E}[\boldsymbol{w}_n] - \overline{\boldsymbol{s}}||_\infty = |\mathbb{E}[w^i_n(j)] - \overline{s}^i(j)| \\
&= |\sigma(s^i_n, j)r(f(\boldsymbol{s}_n), j) + (1-\sigma(s^i_n, j))s^i_n(j) - \overline{s}^i(j)| \\
&= |\sigma(s^i_n, j)(r(f(\boldsymbol{s}_n), j) - \overline{s}^i(j)) + (1-\sigma(s^i_n, j))(s^i_n(j) - \overline{s}^i(j))| \\
&\leqslant |\sigma(s^i_n, j)(r(f(\boldsymbol{s}_n), j) - \overline{s}^i(j))| + |(1-\sigma(s^i_n, j))(s^i_n(j) - \overline{s}^i(j))| \\
&= |\sigma(s^i_n, j)[r(f(\boldsymbol{s}_n), j) - r(f(\overline{\boldsymbol{s}}), j)]| + |(1-\sigma(s^i_n, j))(s^i_n(j) - \overline{s}^i(j))| \\
&\leqslant \sigma(s^i_n, j)C_1||\boldsymbol{s}_n - \overline{\boldsymbol{s}}||_\infty + (1-\sigma(s^i_n, j))|s^i_n(j) - \overline{s}^i(j)| \\
&\leqslant C||\boldsymbol{s}_n - \overline{\boldsymbol{s}}||_\infty.
\end{aligned} \tag{C6}
$$

Here, the second inequality is because $r(f(\boldsymbol{s}_n), j)$ is a contraction, and the last inequality is from the $||\cdot||_\infty$-definition.

The difference in the convergence rate characterizations in Theorem 2 and Corollary 1 is because the state profile $\boldsymbol{s}_t$ is a deterministic process, while the state profile $\boldsymbol{s}_n$ is a stochastic process. Therefore, we use a probabilistic description here for the convergence rate of $\boldsymbol{s}_n$ to the unique MFE $\overline{\boldsymbol{s}}$.

Remark: Though Theorem 2 and Corollary 1 are characterized in line with the asymptotic pseudotrajectory in Lemma 1, convergence results therein are attained mainly using the mapping Γ, instead of directly from off-the-shelf conclusions in stochastic approximation theory.

Appendix C.5 Proof of Theorem 3

Let $\boldsymbol{s}_a$ and $\boldsymbol{s}_b$ be two state profiles, and define a sequence $\boldsymbol{A}_k = [s^1_a, \cdots, s^{k+1}_b, \cdots, s^N_b], k = 1, 2, \cdots, N$ with the first k elements from $\boldsymbol{s}_a$ and the rest from $\boldsymbol{s}_b$. Denote A^i_k as the i-th element (state) of $\boldsymbol{A}_k$, and $\boldsymbol{a}_k$ as the played arms for N agents following the stationary policy $\sigma(A^i_k)$. Let $\boldsymbol{a}^{-i}_k$ be the arm set except agent i, and $f(\boldsymbol{a}_k)$ or $f(\boldsymbol{A}_k)$ be the corresponding population profile. When agent i plays arm j, that is $a^i_k = j$, we have:

$$
\begin{aligned}
|r(f(\boldsymbol{s}_a), j) - r(f(\boldsymbol{s}_b), j)| &= \Big|\sum_{k=1}^N [r(f(\boldsymbol{A}_k), j) - r(f(\boldsymbol{A}_{k-1}), j)]\Big| \\
&\leqslant \sum_{k=1}^N |r(f(\boldsymbol{A}_k), j) - r(f(\boldsymbol{A}_{k-1}), j)| \\
&= \sum_{k=1}^N |r(f(\boldsymbol{a}_k), j) - r(f(\boldsymbol{a}_{k-1}), j)|.
\end{aligned} \tag{C7}
$$

Let $\triangle_k = r(f(\boldsymbol{a}_k), j) - r(f(\boldsymbol{a}_{k-1}), j)$, then $\triangle_i = 0$ since $a^i_k = j$. For $k \neq i$, we express $\triangle_k = \Delta r_k \cdot (\sigma(s^k_a) - \sigma(s^k_b))$, where $\cdot$ is the inner product between two vectors, and Δr_k is a M-length vector with the l-th element $\Delta r_k(l)$ as:

$$
\Delta r_k(l) = \textstyle\sum_{\boldsymbol{z}\in\boldsymbol{a}^{-i}_k, z^k = l} r(f(\{j, \boldsymbol{z}\}), j) \prod_{m\neq k, i} \sigma(A^m_k, z^m). \tag{C8}
$$

Note that both $\sigma(s^k_a)$ and $\sigma(s^k_b)$ are $\triangle^{M-1}$-simplex so that $\Delta r_k(h)\mathbf{1} \cdot (\sigma(s^k_a) - \sigma(s^k_b)) = 0, \forall h \in \mathcal{M}$, where $\mathbf{1}$ is a M-length vector with each element equal to 1. Therefore, we can rewrite $\triangle_k = (\Delta r_k - \Delta r_k(h)\mathbf{1}) \cdot (\sigma(s^k_a) - \sigma(s^k_b))$. If $\boldsymbol{z}_a, \boldsymbol{z}_b \in \boldsymbol{a}^{-i}_k$ with only the k-th arms being different: $z^k_a = l, z^k_b = h$, we obtain $||f(\{j, \boldsymbol{z}_a\}) - f(\{j, \boldsymbol{z}_b\})||_1 \leqslant \frac{2}{N}$ based on the population profile definition. As $r(f_n, j)$ is θ-Lipschitz continuous, we have the result:

$$
|\Delta r_k(l) - \Delta r_k(h)| \leqslant \Big|\sum_{\boldsymbol{z}\in\boldsymbol{a}^{-i}} \prod_{m\neq k, i} \sigma(A^m_k, z^m)\Big| \frac{2\theta}{N} \leqslant \frac{2\theta}{N}. \tag{C9}
$$

Based on Eqs. (C7)-(C9), we obtain:

$$
|r(f(\boldsymbol{s}_a), j) - r(f(\boldsymbol{s}_b), j)| \leqslant \sum_{k=1}^N |\triangle_k| \leqslant \frac{2\theta}{N} \sum_{k=1}^N ||\sigma(s^k_a) - \sigma(s^k_b)||_1. \tag{C10}
$$

We proceed to handle the term $||\sigma(s^k_a) - \sigma(s^k_b)||_1$. Using the mean value theorem, there is a $x \in [s^k_a, s^k_b]$ such that:

$$
\sigma(s^k_a, j) - \sigma(s^k_b, j) = \nabla\sigma(x, j) \cdot (s^k_a - s^k_b). \tag{C11}
$$

Hence, $|\sigma(s_a^k,j)-\sigma(s_b^k,j)| \leqslant ||\nabla\sigma(x,j)||_1||s_a^k-s_b^k||_\infty$ and $||\sigma(s_a^k)-\sigma(s_b^k)||_1 \leqslant \sum_{j=1}^M ||\nabla\sigma(x,j)||_1||s_a^k-s_b^k||_\infty$. Considering the stationary policy of Eq. (1), we compute its derivative as:

$$\frac{d\sigma(x,j)}{dx(l)} = (1-\eta)\beta\sigma(x,j)(\mathbb{1}_{\{j=l\}}-\sigma(x,l)).$$

Therefore:

$$||\nabla\sigma(x,j)||_1 = (1-\eta)\beta\sigma(x,j)\sum_{l=1}^M|(\mathbb{1}_{\{j=l\}}-\sigma(x,l))| = 2(1-\eta)\beta\sigma(x,j)(1-\sigma(x,j)) \leqslant 2(1-\eta)\beta\sigma(x,j). \tag{C12}$$

Combining with Eqs. (C11) and (C12), we obtain:

$$||\sigma(s_a^k,j)-\sigma(s_b^k,j)||_1 \leqslant \sum_{j=1}^M 2(1-\eta)\beta\sigma(x,j)||s_a^k-s_b^k||_\infty \leqslant 2(1-\eta)\beta||\boldsymbol{s}_a-\boldsymbol{s}_b||_\infty. \tag{C13}$$

In line with Eq. (C10), we present the final result:

$$|r(f(\boldsymbol{s}_a),j)-r(f(\boldsymbol{s}_b),j)| \leqslant 4\theta(1-\eta)\beta||\boldsymbol{s}_a-\boldsymbol{s}_b||_\infty. \tag{C14}$$

Appendix C.6 Contraction mapping for linear reward

The condition $4\theta(1-\eta)\beta<1$ in Theorem 3 is a little stringent due to twofold reasons. First, the reward function $r(f(\boldsymbol{s}_t),j)$ depends on *all* M *elements* of the population profile $f(\boldsymbol{s}_t)$. Second, $r(f(\boldsymbol{s}_t),j)$ is *non-linear* in $f(\boldsymbol{s}_t)$ so that calculating the expected reward needs multiple scaling operations. To relax the condition, we make two reasonable assumptions. In fact, the reward $r(f(\boldsymbol{s}_t),j)$ of playing arm j is often only impacted by the number of agents who select arm j [8]. To name a few, we take the resource competition game in Appendix B, the first three applications, as an example, where agents claim for one of the M types of resource (arms) in each time slot. An agent playing arm j only competes with those making the same choice. Hence, we assume $r(f(\boldsymbol{s}_t),j)$ merely depends on the j-th element of the population profile, denoted as $f(\boldsymbol{s}_t,j)$. Besides, we further presume the reward is a linear function in the population profile. Hence, one can move the expectation when computing the reward directly inside on the population profile. With this two assumptions, we recharacterize a relaxed condition for the contraction mapping.

Corollary 2. If the reward $r(f(\boldsymbol{s}_t,j),j)$ is a θ-Lipschitz continuous linear function in the j-th element $f(\boldsymbol{s}_t,j)$, then $r(f(\boldsymbol{s}_t,j),j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$ under the condition $\frac{\theta(1-\eta)\beta}{2}<1$, where β,η are from Eq. (1).

Proof. Due to the linearity, we move the expectation into on the population profile. From Eqs. (1)-(2), the expected population profile, with a slight abuse of notations, is:

$$f(\boldsymbol{s}_t,j) = \mathbb{E}\Big[\sum_{i=1}^N\frac{\mathbb{1}_{\{a_t^i=j\}}}{N}\Big] = \frac{\sum_{i=1}^N\mathbb{E}\Big[\mathbb{1}_{\{a_t^i=j\}}\Big]}{N} = \frac{\sum_{i=1}^N\sigma(s_t^i,j)}{N}. \tag{C15}$$

Since the reward function $r(f(\boldsymbol{s}_t,j),j)$ is θ-Lipschitz continuous in the population profile $f(\boldsymbol{s}_t,j)$, we obtain $|\frac{dr(f(\boldsymbol{s}_t,j),j)}{df(\boldsymbol{s}_t,j)}| \leqslant \theta$. Applying the mean value theorem, there is a vector $\boldsymbol{x}\in[\boldsymbol{s}_a,\boldsymbol{s}_b]$ satisfying the condition:

$$r(f(\boldsymbol{s}_a,j),j)-r(f(\boldsymbol{s}_b,j),j) = \nabla r(f(\boldsymbol{x},j),j)\cdot(\boldsymbol{s}_a-\boldsymbol{s}_b).$$

Therefore, $|r(f(\boldsymbol{s}_a,j),j)-r(f(\boldsymbol{s}_b,j),j)| \leqslant ||\nabla r(f(\boldsymbol{x},j),j)||_1||\boldsymbol{s}_a-\boldsymbol{s}_b||_\infty$. Our main work is left to bound the derivative $\nabla r(f(\boldsymbol{x},j),j)$.

Utilizing the chain of derivative, we have $\frac{dr(f(\boldsymbol{x},j),j)}{dx^i(l)} = \frac{dr(f(\boldsymbol{x},j),j)}{df(\boldsymbol{x},j)}\times\frac{df(\boldsymbol{x},j)}{dx^i(l)}, \forall i\in\mathcal{N}, l\in\mathcal{M}$. Since $|\frac{dr(f(\boldsymbol{x},j),j)}{df(\boldsymbol{x},j)}| \leqslant \theta$, we only need to handle the second term $\frac{df(\boldsymbol{x},j)}{dx^i(l)}$. Recall from the stationary policy of Eq. (1) and the expected population profile in Eq. (C15), we obtain that:

$$\frac{df(\boldsymbol{x},j)}{dx^i(l)} = \frac{1}{N}\frac{d\sigma(x^i,j)}{dx^i(l)} = \frac{(1-\eta)\beta\sigma(x^i,j)(\mathbb{1}_{\{j=l\}}-\sigma(x^i,l))}{N}.$$

Hence, we have:

$$\begin{aligned}
||\nabla r(f(\boldsymbol{x},j),j)||_1 &= \sum_{i=1}^N\sum_{l=1}^M\left|\frac{dr(f(\boldsymbol{x},j),j)}{dx^i(l)}\right| \\
&\leqslant \frac{\theta(1-\eta)\beta}{N}\sum_{i=1}^N\sum_{l=1}^M|\sigma(x^i,j)(\mathbb{1}_{\{j=l\}}-\sigma(x^i,l))| \\
&= \frac{\theta(1-\eta)\beta}{N}\sum_{i=1}^N 2\sigma(x^i,j)(1-\sigma(x^i,j)) \\
&\leqslant \frac{\theta(1-\eta)\beta}{N}\sum_{i=1}^N 2\times\frac{1}{4} \\
&= \frac{\theta(1-\eta)\beta}{2},
\end{aligned} \tag{C16}$$

where the second inequality is because $y(1-y)\leqslant\frac{1}{4}, \forall y\in[0,1]$. Finally, it results in:

$$|r(f(\boldsymbol{s}_a,j),j)-r(f(\boldsymbol{s}_b,j),j)| \leqslant \frac{\theta(1-\eta)\beta}{2}||\boldsymbol{s}_a-\boldsymbol{s}_b||_\infty. \tag{C17}$$

Comparing the conditions in Theorem 3 and Corollary 2, the factor 4 is now reduced to $\frac{1}{2}$. Therefore, parameters θ,β,η can take much broader range of values to ensure the contraction mapping.

Appendix D State change and model extension

Appendix D.1 Proof of Theorem 4

Appendix D.1.1 *Regret description*

The bandit game will not converge to an equilibrium when applying the traditional MAB algorithms which may have tight-bounded regret, like UCB and EXP3. Because they need to model an agent state as the cumulative reward to determine the playing policy. Hence, any state will consistently "increase" to "explosion" since realized rewards are positive values, i.e., the system is unstable. This is also the underlying reason why [8–10] need to assume the state regeneration to reset agents' state with certain probability so as to obtain the mean field equilibrium. Actually, traditional MAB algorithms, such as UCB and its variants, have sublinear regrets only in single-agent or stable multi-agent scenarios, while their regrets can not be guaranteed in unstable multi-agent systems. We will also elaborate this point in the performance evaluation.

Appendix D.1.2 *Main proof*

Let $X_n^i(j) = \text{Exp}(\beta s_n^i(j))$ and $X_n^i = \sum_{j=1}^M X_n^i(j)$. Combining with Eqs. (1) and (2), we obtain:

$$\begin{aligned}
\frac{X_{n+1}^i}{X_n^i} &= \sum_{j=1}^M \frac{X_{n+1}(j)}{X_n} \\
&= \sum_{j=1}^M \frac{X_n^i(j)}{X_n^i} \text{Exp}[\beta\gamma_n(w_n^i(j) - s_n^i(j))] \\
&= \sum_{j=1}^M \frac{\sigma(s_n^i, j) - \frac{\eta}{M}}{1-\eta} \text{Exp}[\beta\gamma_n(w_n^i(j) - s_n^i(j))] \qquad \text{(D1)} \\
&\leqslant \sum_{j=1}^M \frac{\sigma(s_n^i, j) - \frac{\eta}{M}}{1-\eta} [1 + \beta\gamma_n(w_n^i(j) - s_n^i(j)) + (e-2)\beta^2\gamma_n^2(w_n^i(j) - s_n^i(j))^2] \\
&\leqslant 1 + \sum_{j=1}^M \frac{\sigma(s_n^i, j) - \frac{\eta}{M}}{1-\eta} \beta\gamma_n(w_n^i(j) - s_n^i(j)) + \sum_{j=1}^M \frac{(e-2)\sigma(s_n^i, j)\beta^2\gamma_n^2}{1-\eta}(w_n^i(j) - s_n^i(j))^2,
\end{aligned}$$

where the first inequality is because $\text{Exp}(y) \leqslant 1 + y + (e-2)y^2$ and e is the Euler's number. Using the result $\ln y \leqslant y - 1, \forall y > 0$ and the fact that $\ln \frac{X_{T+1}^i}{X_0^i} = \sum_{n=0}^T \ln \frac{X_{n+1}^i}{X_n^i}$, we have:

$$\ln \frac{X_{T+1}^i}{X_0^i} \leqslant \sum_{n=0}^T \sum_{j=1}^M \frac{\sigma(s_n^i, j) - \frac{\eta}{M}}{1-\eta} \beta\gamma_n(w_n^i(j) - s_n^i(j)) + \sum_{n=0}^T \sum_{j=1}^M \frac{(e-2)\sigma(s_n^i, j)\beta^2\gamma_n^2}{1-\eta}(w_n^i(j) - s_n^i(j))^2. \qquad \text{(D2)}$$

For any arm j, it satisfies:

$$\ln \frac{X_{T+1}^i}{X_0^i} \geqslant \ln \frac{X_{T+1}^i(j)}{X_0^i}. \qquad \text{(D3)}$$

Comparing Eqs. (D2) and (D3), we obtain:

$$\ln \frac{X_{T+1}^i(j)}{X_0^i} \leqslant \sum_{n=0}^K \sum_{j=1}^M \frac{\sigma(s_n^i, j) - \frac{\eta}{M}}{1-\eta} \beta\gamma_n(w_n^i(j) - s_n^i(j)) + \sum_{n=0}^T \sum_{j=1}^M \frac{(e-2)\sigma(s_n^i, j)\beta^2\gamma_n^2}{1-\eta}(w_n^i(j) - s_n^i(j))^2. \qquad \text{(D4)}$$

Moreover, $\ln \frac{X_{T+1}^i(j)}{X_0^i} = \beta s_0^i(j) + \sum_{n=0}^T \beta\Delta s_n^i(j) - \ln(\sum_{j=1}^M \text{Exp}(\beta s_0^i(j)))$. Using the expression of the inner product between two vectors, we complete the proof.

Appendix D.2 Heterogeneous learning parameter

So far, we have analyzed the bandit game when the stationary policy adopts homogeneous learning parameters, that is uniform β, η in Eq. (1). Next, we explore the heterogeneous situation where β, η could vary for different agents.

For agent i, parameter β^i will keep unchanged over time. In contrast, value of η_n^i is diminishing and satisfies $\lim_{n\to\infty} \eta_n^i = 0$. The reason is to give less weight to the random choice when the reward information is accurately learned. Therefore, the stationary policy for the heterogeneous case is:

$$\sigma(s_n^i, j) = (1 - \eta_n^i) \frac{\text{Exp}\left(\beta^i s_n^i(j)\right)}{\sum_{k=1}^M \text{Exp}\left(\beta^i s_n^i(k)\right)} + \frac{\eta_n^i}{M}. \qquad \text{(D5)}$$

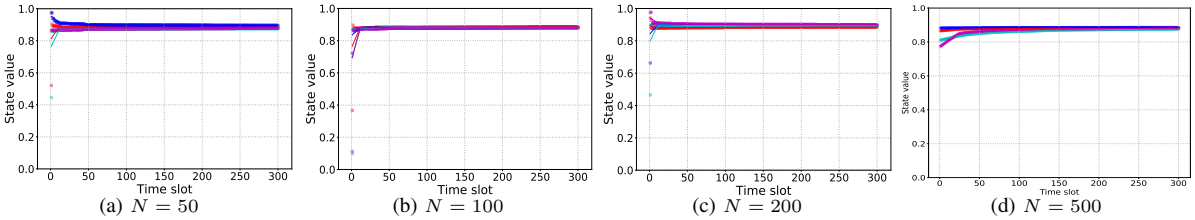
Except the playing policy, the updating rule is the same as the homogeneous case, described in Eq. (2). In the following, we illustrate the convergence property and the cumulative state change when adopting heterogeneous parameters. We will only present the results while omitting the proofs, since they share a similar spirit of the homogeneous situation.

Corollary 3. If agents play actions in accordance to Eq. (D5), the following conclusions stand.

1) There exists an MFE $\overline{\boldsymbol{s}}$ such that $\overline{\boldsymbol{s}} \in \Gamma(\overline{\boldsymbol{s}})$.

2) For the state s_n^i, its interpolated process $\tilde{s}_t^i$ is an asymptotic pseudotrajectory for the ODE defined in Lemma 1.

3) If the reward function $r(f(\boldsymbol{s}_t), j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$, there is a unique MFE $\overline{\boldsymbol{s}}$ for the bandit game, which is also a global attractor for the ODE, and $\boldsymbol{s}_t$ converges to $\overline{\boldsymbol{s}}$ with exponential rate. Convergence rate of the state profile $\boldsymbol{s}_n$ in Corollary 1 also holds.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E1 State under contraction general reward.

4) Assume the reward function $r(f_n, j)$ is θ-Lipschitz continuous in the population profile f_n, and denote $\beta_{\max} = \max\{\beta^i | i \in \mathcal{N}\}$. If $4\theta\beta_{\max} < 1$, then $r(f(\boldsymbol{s}_t), j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$. If $r(f_n(j), j)$ is linear in the j-th element $f_n(j)$, the condition is relaxed to $\frac{\theta\beta_{\max}}{2} < 1$.

5) $\forall i \in \mathcal{N}, \forall j \in \mathcal{M}$, the cumulative state change satisfies:

$$\beta^i s_0^i(j) + \sum_{n=0}^{T} \beta^i \Delta s_n^i(j) - \ln\Big(\sum_{j=1}^{M} \mathrm{Exp}\left(\beta^i s_0^i(j)\right)\Big) \leqslant \sum_{n=0}^{T} \Big[\frac{\beta^i(\sigma(s_n^i) - \frac{\eta_n^i}{M}\mathbf{1}) \cdot \Delta s_n^i}{1-\eta_n^i} + \frac{(e-2)(\beta^i)^2\sigma(s_n^i)\cdot(\Delta s_n^i)^2}{1-\eta_n^i}\Big]. \tag{D6}$$

Appendix D.3 Overlapping arms

Another extension is about overlapping arms. In Appendix B, we provide several applications in the computer communication and services computing area. Sometimes, an agent can only get access to a subset of arms, such as a mobile user could merely offload tasks to a subset of edge servers due to the servers' limited radio coverage. Mathematically, there are M arms, and agent i pulls an arm from a subset $\mathcal{M}^i, \mathcal{M}^i \subseteq \mathcal{M}$ with the cardinality being M^i. Overlapping means there exist agents i, k such that $\mathcal{M}^i \cap \mathcal{M}^k \neq \emptyset$. The conclusion is: previous results still hold after we make the following main adjustments. The proofs are the same, so that we skip them to save space.

One adjustment is to choose arms from $\mathcal{M}^i$ for agent i:

$$\sigma(s_n^i, j) = (1-\eta)\frac{\mathrm{Exp}\left(\beta s_n^i(j)\right)}{\sum_{k\in\mathcal{M}^i}\mathrm{Exp}\left(\beta s_n^i(k)\right)} + \frac{\eta}{M^i}. \tag{D7}$$

The state s_n^i for agent i is now a M^i-length vector, and the state profile $\boldsymbol{s}_n$ is in $[0,1]^{\sum_{i\in\mathcal{N}} M^i}$. Other parameters, especially the cardinality of sets, are adapted accordingly. Following the same approach of the non-overlapping case, one can obtain the existence and uniqueness of MFE.

Appendix D.4 Results for logit policy

Previous contents are devoted to deriving the existence and uniqueness of MFE under the Hedge stationary policy. We now extend the mean field model to consider the case where agents follow the logit policy, namely the probability of choosing arm j is:

$$\sigma(s_n^i, j) = \frac{\mathrm{Exp}\left(\beta s_n^i(j)\right)}{\sum_{k=1}^{M}\mathrm{Exp}\left(\beta s_n^i(k)\right)}. \tag{D8}$$

Aside from the playing policy, the updating rule of Eq. (2) maintains unchanged. Following a similar approach of the MFE characterization, we acquire the results, slightly different to the Hedge stationary policy, as below.

Corollary 4. If agents follow the logit policy of Eq. (D8), we draw the following conclusions.

1) There exists an MFE $\overline{\boldsymbol{s}}$ such that $\overline{\boldsymbol{s}} \in \Gamma(\overline{\boldsymbol{s}})$.

2) For state s_n^i, its interpolated process $\tilde{s}_t^i$ is an asymptotic pseudotrajectory for the ODE in Lemma 1.

3) If the reward function $r(f(\boldsymbol{s}_t), j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$, there is a unique MFE $\overline{\boldsymbol{s}}$ for the bandit game, which is also a global attractor for the ODE, and $\boldsymbol{s}_t$ converges to $\overline{\boldsymbol{s}}$ with exponential rate. Convergence rate of the state profile $\boldsymbol{s}_n$ in Corollary 1 also holds.

4) Assume the reward function $r(f_n, j)$ is θ-Lipschitz continuous in the population profile f_n. If $4\theta\beta < 1$, then $r(f(\boldsymbol{s}_t), j)$ is a $||\cdot||_\infty$-contraction in the state profile $\boldsymbol{s}_t$. If $r(f_n(j), j)$ is linear in the j-th element $f_n(j)$, the condition is relaxed to $\frac{\theta\beta}{2} < 1$.

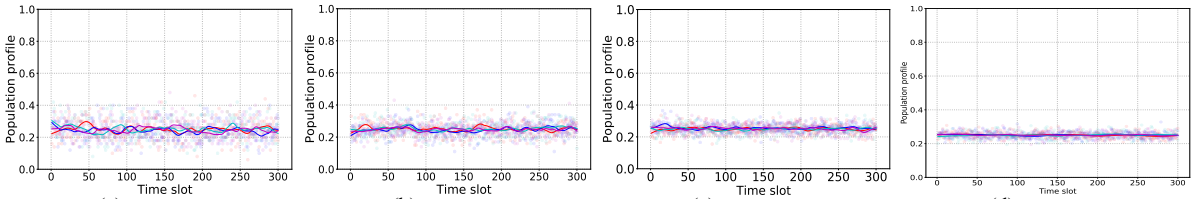
5) $\forall i \in \mathcal{N}, \forall j \in \mathcal{M}$, the cumulative state change satisfies:

$$\beta s_0^i(j) + \sum_{n=0}^{T} \beta \Delta s_n^i(j) - \ln\Big(\sum_{j=1}^{M} \mathrm{Exp}\left(\beta s_0^i(j)\right)\Big) \leqslant \sum_{n=0}^{T} \left[\beta\sigma(s_n^i)\cdot\Delta s_n^i + (e-2)\beta^2\sigma(s_n^i)\cdot(\Delta s_n^i)^2\right]. \tag{D9}$$

The proof for Corollary 4 is omitted since it is akin to the Hedge stationary policy. This corollary implies the mean field model can be applied to a series of arm playing policies, such as the multiplicative weights update methods [11].

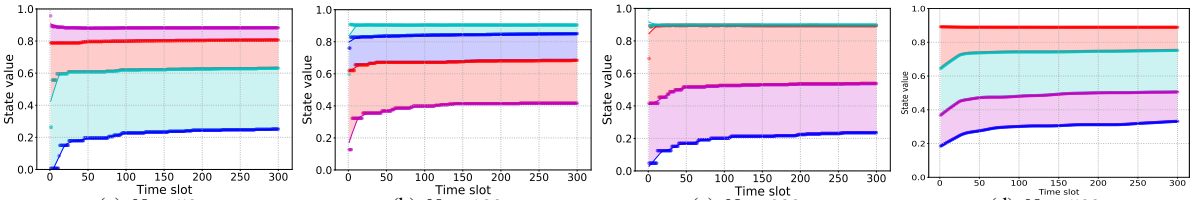
Appendix E Performance evaluation

In this section, we evaluate the performance of the bandit game. We investigate the situation where agents' actions have negative externalities on rewards, i.e., the reward decreases as the number of agents simultaneously choosing the same arm increases. As the existence and uniqueness of MFE depend on the reward function, both general and linear reward forms will be implemented. Moreover, the corresponding results are smoothed by *lowess* in Python for better exhibition.



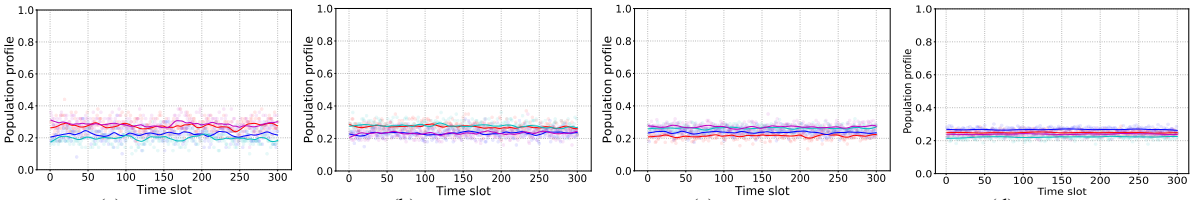
(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E2 Population profile under contraction general reward.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E3 State under non-contraction general reward.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E4 Population profile under non-contraction general reward.

Appendix E.1 General reward function

Reward function. Like the applications enumerated in Appendix B, we consider that the general reward $r(f_n, j)$ is a non-linear function in the j-th element of the population profile, or $f_n(j)$, where $f_n(j)$ will be similarly in the denominator part of $r(f_n, j)$. With this in mind, we use a concise while typical reward function to perform the evaluations [8]:

$$r(f_n, j) = \frac{1}{1 + \theta(j) f_n(j)}, \tag{E1}$$

where $\theta(j) \in [0.8\theta, \theta], \forall j \in \mathcal{M}$. One can verify that $r(f, j)$ is in $[0, 1]$ and satisfies θ-Lipschitz continuity. Besides, the stepsize is $\gamma_n = \frac{1}{n+1}$.

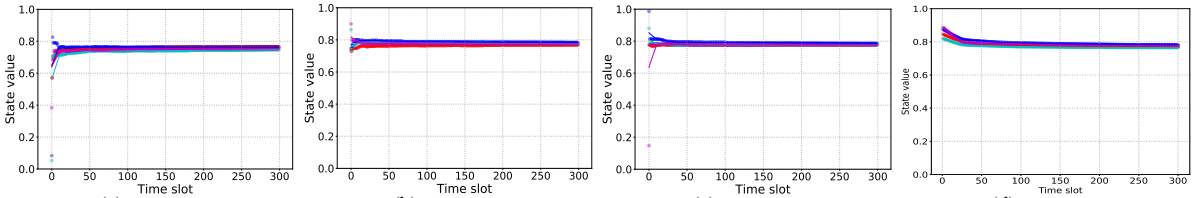
Performance under contraction mapping. We first show the results under the $||\cdot||_\infty$-contraction mapping. From Theorem 3, parameters (θ, β, η) are set to $(0.5, 0.5, 0.2)$, respectively, and hence the contraction condition $4\theta(1-\eta)\beta < 1$ holds. Each state is initialized to be a random value in $[0, 1]$. Given $M = 4$, i.e., the number of arms is 4, we run the bandit game for four times with each lasting for 300 time slots, where the state is reinitialized and other parameters keep unchanged in each run. In Figure E1, we display the state evolution of arm 2 when the agent number N varies. We can see that the state will converge to a fixed value, which is also unique in different runs for each N. Hence, a unique MFE is obtained if the reward function is a contraction mapping. Notably, traditional Markov game theory grapples with managing a state space of $\mathcal{S}^N$, where $\mathcal{S}$ represents the state of a single agent. This becomes computationally prohibitive when managing large-scale systems, as demonstrated with $N = 200$ or $N = 500$, rendering such approaches impractical for extensive agent populations.

We further demonstrate the population profile evolution, and plot its trend over time in Figure E2. We observe that the population profile of arm 2 is unique, and tends to be stable as N increases. In other words, the fluctuation of the population profile becomes smaller when N is larger. This can be explained through the Chebyshev's inequality. Suppose the unique equilibrium is $\overline{\boldsymbol{s}}$, then the expected population profile $\mathbb{E}[f(\overline{\boldsymbol{s}}, j)]$ is $\frac{\sum_{i=1}^N \sigma(\overline{s}^i, j)}{N}$. Since agent i chooses arm j with probability $\sigma(\overline{s}^i, j)$, we have $\mathbb{E}\left[\mathbb{1}_{\{\overline{a}^i = j\}}\right] = \sigma(\overline{s}^i, j)$, and the variance $\text{var}\left[\mathbb{1}_{\{\overline{a}^i = j\}}\right] = \sigma(\overline{s}^i, j)(1 - \sigma(\overline{s}^i, j)) \leqslant \frac{1}{4}$ due to $y(1-y) \leqslant \frac{1}{4}, \forall y \in [0, 1]$. As agents play arms independently, then $\text{var}[f(\overline{\boldsymbol{s}}, j)] = \frac{1}{N^2} \sum_{i=1}^N \text{var}\left[\mathbb{1}_{\{\overline{a}^i = j\}}\right] \leqslant \frac{1}{N^2} \frac{N}{4} = \frac{1}{4N}$. Based on the Chebyshev's inequality, we obtain:

$$\Pr\left(|f(\overline{\boldsymbol{s}}, j) - \mathbb{E}[f(\overline{\boldsymbol{s}}, j)]| \geqslant \epsilon\right) \leqslant \frac{1}{4N\epsilon^2}. \tag{E2}$$

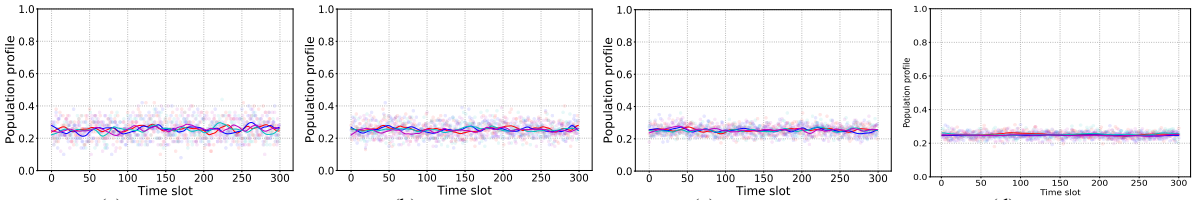
Therefore, if N increases, the empirical $f(\overline{\boldsymbol{s}}, j)$ will deviate from $\mathbb{E}[f(\overline{\boldsymbol{s}}, j)]$ with lower probability.

Performance under non-contraction mapping. We continue studying the case where the contraction mapping is violated. In particular, parameters θ, η and M stay the same as those in the former case, while β changes to 30 so that the contraction mapping condition $4\theta(1-\eta)\beta < 1$ does not stand. Similarly, we run the bandit game for 300 time slots to exhibit the trends clearly. For each run, we repeat the simulation for four times by reassigning initial agent states in $[0, 1]$. The state evolution is depicted in Figure E3, which shows that the state converges to multiple distinct fixed points, namely different MFEs. Moreover, we plot the population profile evolution in Figure E4. Comparing with Figure E2, we can also observe that the fluctuation around $\mathbb{E}[f(\overline{\boldsymbol{s}}, j)]$



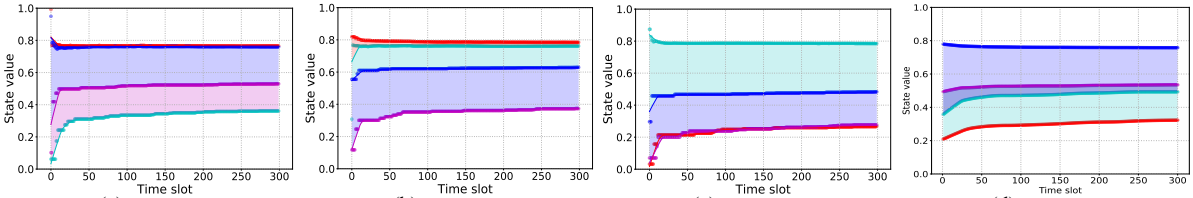
(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E5 State under contraction linear reward.



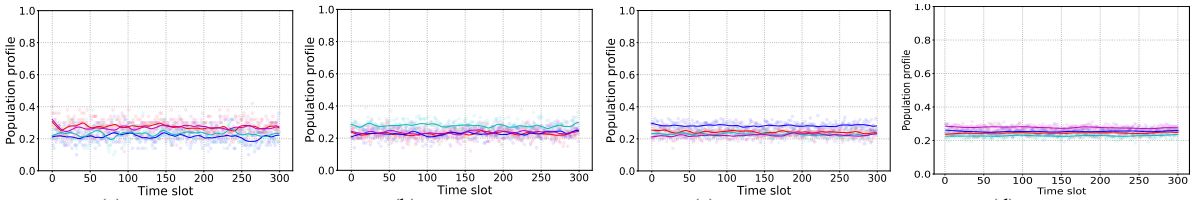
(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E6 Population profile under contraction linear reward.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E7 State under non-contraction linear reward.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$ (d) $N = 500$

Figure E8 Population profile under non-contraction linear reward.

becomes impaired when N is large. Since there are multiple MFEs, the population profile will converge to different steady values as well.

Appendix E.2 Linear reward function

Reward function. In Eq. (E1), we instantiate a non-linear function to represent the general reward. Given that the contraction mapping condition (see Corollary 2) for the linear reward function is less stringent, we proceed to further evaluate the mean field equilibrium (MFE) under linear rewards. The scenario with linear rewards can be considered a special case of the general reward, distinct from the example provided in Eq. (E1). In line with the negative externality, reward $r(f_n(j), j)$ has the following expression:

$$r(f_n(j), j) = 1 - \theta(j) f_n(j), \tag{E3}$$

where $\theta(j) \in [0.8\theta, \theta], \forall j \in \mathcal{M}$ with $\theta \in [0, 1]$. It can be identified that $r(f_n(j), j)$ falls into the range $[0, 1]$, and maintains θ-Lipschitz continuity as well.

Performance under contraction mapping. Similarly, we illustrate the results when $||\cdot||_\infty$-contraction mapping stands in the first place. According to Corollary 2, values of (θ, β, η) are assigned to $(1, 2, 0.2)$, respectively, thereby meeting the contraction condition $\frac{\theta(1-\eta)\beta}{2} < 1$. The arm number $M = 4$, and each state is initialized to be a value in $[0, 1]$. Still let the simulation proceeds for four runs, with each run lasting for 300 time slots and states being reinitialized. The state evolution of a selected agent's arm 2 is depicted in Figure E5. We can see that the curves will approach a specific value for different agent number N, i.e., unique MFE is derived due to contraction mapping of the reward function. As for the population profile shown in Figure E6, we also obtain that it tends to be more stable around a unique value when N increases, which can be similarly explained by the Chebyshev's inequality in Eq. (E2).

Performance under non-contraction mapping. Now we evaluate the performance of linear reward function when the contraction mapping condition is no longer satisfied. In particular, parameters θ, η and M remain unchanged, while β is altered to 40. The simulation runs for four times, and each time the initial state value is refreshed. We show the results for state evolution

Table E1 Empirical regret under contraction/non-contraction reward.

Reward function	Term	$N = 50$		$N = 100$		$N = 200$		$N = 500$	
		Contraction	Non-contraction	Contraction	Non-contraction	Contraction	Non-contraction	Contraction	Non-contraction
General	Regret	14.653	21.187	13.758	19.014	7.787	12.108	11.693	17.006
	Reward	1791.904	1786.238	1796.961	1784.582	1791.061	1791.829	1748.343	1743.035
Linear	Regret	21.464	28.362	19.023	26.626	17.932	23.947	22.606	25.728
	Reward	1541.083	1552.282	1554.948	1558.675	1560.613	1559.661	1496.204	1502.768

Table E2 Empirical regret of UCB.

Reward function	Term	$N = 50$	$N = 100$	$N = 500$	$N = 200$
General	Regret	58.194	58.998	43.149	60.682
	Reward	1779.195	1769.988	1779.495	1756.591
Linear	Regret	63.285	80.627	41.665	59.970
	Reward	1537.043	1565.712	1575.706	1487.256

and population profile evolution in Figures E7 and E8, respectively. Like the general reward, the state will reach different steady values, which amount to multiple MFEs. Moreover, the population profile also converges to varied fixed values with the fluctuation being impaired for larger agent number N.

Appendix E.3 Empirical regret

The cumulative state change is derived in Theorem 4 which can only loosely bound the theoretical scaled regret. However, we will show that the stationary policy in fact has a tight empirical regret, even compared with EXP3 [12]. For both the general reward function in Eq. (E1) and the linear reward function in Eq. (E3), we compute the regrets when parameters maintain unchanged for the contraction and non-contraction reward function.

Regret under contraction mapping. We conduct the simulation for six times with each running for $T = 2000$ time slots, and illustrate the average regret and cumulative reward in Table E1. A well-known regret bound for EXP3 is $O(\sqrt{T})$, and here $\sqrt{T} = 44.721$. Regarding the general reward function, the regret is much smaller than $\sqrt{T}$. As for the linear reward function, the regret is also very small, which is especially conspicuous when N increases. Moreover, we have two interesting observations. First, the regret of the general reward is less than that of the linear reward. This is because changes in the population profile have smaller impact on the general reward since it appears in the denominator of Eq. (E1). Second, the regret deceases as N grows large for both cases. The main reason is the population profile becomes more stable, thus the gap between the maximum and realized rewards is reduced.

Regret under non-contraction mapping. We further provide the result if the general/linear reward function does not meet the contraction mapping condition. Compared with the contraction case, an agent state may diverge to various MFEs. The simulation repeats itself for six times, where each run also lasts for $T = 2000$ time slots in total. Both the cumulative reward and regret for the two reward functions are exhibited in Table E1. For the cumulative reward, both general function and linear function yield similar results to those for the contraction case. Nevertheless, the regrets have different behaviors. Pertaining to non-contraction reward functions, the corresponding regrets have large variances, with the average values bigger than those for the contraction case. One of the main reasons behind is that there exist multiple MFEs, which will cause varied and large (on average) regrets. But at the same time, the regrets for general and linear functions are still less than the bound $\sqrt{T} = 44.721$. Besides, regrets tend to decrease as agent number N increases large, and the general function has smaller regret than the linear function as well. To sum up, the stationary policy still has tight empirical regrets for both contraction and non-contraction reward functions.

Regret of UCB. For comparisons, we present the regret when replacing the stationary policy with UCB [13]. As aforementioned, the regret bound for UCB is unavailable since it will lead to an unstable state. To show this, we proceed the evaluations for six times, and provide the average regrets and cumulative rewards in Table E2 for general and linear rewards. Results verify our previous claim, i.e., the regrets of UCB for both reward functions are very large due to the unstable state. Compared to the stationary policy even under non-contraction mappings, the regret of UCB are still much higher. Therefore, obtaining MFE is of paramount importance in that a guaranteed system performance including regret is attained when MFE could be derived.

References

1 Samarakoon S, Bennis M, Saad W, et al. Ultra Dense Small Cell Networks: Turning Density into Energy Efficiency. IEEE JSAC, 2016, 34: 1256-1280.
2 Hanif A F, Tembine H, Assaad M, et al. Mean-Field Games for Resource Sharing in Cloud-Based Networks. IEEE/ACM TON, 2016, 24: 624-637.
3 Kunreuther H, Heal G. Interdependent Security. Journal of Risk and Uncertainty, 2003, 26: 231-249.
4 Adlakha S, Johari R. Mean Field Equilibrium in Dynamic Games with Strategic Complementarities. Operations Research, 2013, 61: 971-989.
5 Lasry J, Lions P. Mean field games. JJM, 2007, 2: 229–260.
6 Benaim M. Dynamics of stochastic approximation algorithms. Seminaire de probabilites, 1999, 1-68.
7 Dar E E, Mansour Y. Learning Rates for Q-learning. JMLR, 2003, 1-25.
8 Gummadi R, Johari R, Schmit S, et al. Mean Field Analysis of Multi-Armed Bandit Games. [Online]. Avaliable: https://ssrn.com/abstract=2045842, 2013.
9 Maghsudi S, Hossain E. Distributed User Association in Energy Harvesting Dense Small Cell Networks: A Mean-Field Multi-Armed Bandit Approach. IEEE Access, 2017, 5: 3513-3523.
10 Zhao Z, Liu A L. Intelligent Demand Response for Electricity Consumers: A Multi-armed Bandit Game Approach. In: Proceedings of IEEE ISAP, San Antonio, TX, USA, 2017. 1-6.
11 Arora S, Hazan E, Kale S. The Multiplicative Weights Update Method: A Meta-Algorithm and Applications. Theory of Computing, 2012, 8: 121-164.
12 Auer P, Cesa B N, Freund Y, et al. The Nonstochastic Multiarmed Bandit Problem. SIAM J. Comput., 2002, 32: 2002.
13 Auer P, Cesa B N, Fischer P S. Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning, 2002, 47: 235-256.