Mean Field Analysis of Multi-Armed Bandit Games

Ramki Gummadi
Department of Management Science and Engineering, Stanford University; gummadi@stanford.edu

Ramesh Johari
Department of Management Science and Engineering, Stanford University; rjohari@stanford.edu

Sven Schmit
Institute for Computational and Mathematical Engineering, Stanford University; schmit@stanford.edu

Jia Yuan Yu
IBM Research, Dublin; jiayuanyu@ie.ibm.com

Much of the classical work on algorithms for multi-armed bandits focuses on rewards that are stationary over time. By contrast, we study *multi-armed bandit (MAB) games*, where the rewards obtained by an agent also depend on how many other agents choose the same arm (as might be the case in many competitive or cooperative scenarios). Such systems are naturally nonstationary due to the interdependent evolution of agents, and in general MAB games can be intractable to analyze using typical equilibrium concepts (such as perfect Bayesian equilibrium).

We introduce a general model of multi-armed bandit games, and study the dynamics of these games under a large system approximation. We investigate conditions under which the bandit dynamics have a steady state we refer to as a *mean field steady state* (MFSS). In an MFSS, the proportion of agents playing the various arms, called the *population profile*, is assumed stationary over time; the steady state definition then requires a consistency check that this stationary profile arises from the policies chosen by the agents.

We establish the following results in the paper. First, we establish existence of an MFSS under broad conditions. Second, we show under a *contraction* condition that the MFSS is unique, and that the population profile converges to it from any initial state. Finally, we show that under the contraction condition, MFSS is a good approximation to the behavior of finite systems with many agents. The contraction condition requires that the agent population regenerates sufficiently often, and that the sensitivity of the reward function to the population profile is low enough. Through numerical experiments, we find that in settings with negative externalities among the agents, convergence obtains even when our condition is violated; while in settings with positive externalities among the agents, our condition is tighter.

Key words: Multi-armed Bandits; Mean field equilibria; Markov chains

---

1. Introduction Multi-armed bandit (MAB) models are a key framework for representing sequential decision-making under uncertainty. In classical formulations of the MAB problem, an agent can choose among a finite set of alternatives (arms). At each time period, the agent makes a single choice (a "pull"), and earns a (random) reward. At the outset, the agent may be uncertain about the distribution of the reward she will obtain by pulling any given arm; this information is only learned by experimentation. Thus the agent faces the well-studied trade-off between *exploration*—learning more about alternatives not yet fully explored; and *exploitation*—taking advantage of alternatives that have already proven to be profitable.

The MAB model captures a diverse range of scenarios where agents learn from their actions. In many settings, it is commonly assumed that the environment in which the agent lives is *stationary*; in the model of the previous paragraph, this means that the reward distributions, while unknown, are fixed over the lifetime of the agent. Of course, in practice, there are many potential sources of nonstationarity, sometimes due simply to exogenous evolution of the agent's environment.

In this paper, we focus on a particular source of nonstationarity in MAB settings: *interaction with other agents*. In many cases, an agent appears to be solving an MAB problem, but in fact the rewards earned on the arms may be highly dependent on the actions of *other* agents who are also

solving their own MAB problems. For example, in wireless resource sharing, when a device chooses to transmit in a given channel, it competes directly with other devices exploring the same channel. In most models of competition, if a firm tries to sell a new product, or enter a new market, it directly competes against other firms already selling that product or in that market. Similarly, in online ad auctions, advertisers considering bidding on different keywords must consider the likelihood of winning against other bidders competing on the same keywords.

When agents interact in this way, the overall system can no longer be analyzed through the eyes of a single agent; rather, we view the agents' interactions as a dynamic game, that we call a *multi-armed bandit (MAB) game*. Somewhat surprisingly, while MAB problems and variants have been extensively studied, there is little structural insight into dynamic games where agents solve interlinked MAB problems. In this paper, we provide significant insight into this class of strategic models.

To illustrate the difficulty, first suppose that two agents each choose among a finite set of arms each period, and that the (random) rewards they obtain are dependent on both an agent specific parameter, as well as whether or not the other agent pulled the same arm. It should be clear that this problem is significantly more complex than the classical MAB, because by pulling one arm, an agent learns both about that arm's reward distribution *and* her opponent's strategy. For example, in a resource sharing scenario, if an agent does not receive a reward on an arm, this would likely increase her belief that another player may have been taking the same action at the same time.

Of course, a proper analysis of dynamic equilibrium in such a game requires modeling each player's beliefs, their beliefs over other players beliefs, etc. Indeed, in such a dynamic game with finitely many players, the standard equilibrium concept is perfect Bayesian equilibrium (PBE); PBE requires that: (1) agents maintain beliefs over all that they learn about their competitors; and (2) agents play optimally after any history, given their beliefs. The resulting equilibrium concept is both *intractable* (as it requires exceedingly complex state information) and *implausible* (since in practice, agents may not track fine-scale behavior of their competitors).

The above pitfalls draw attention back to the fundamental motivation for introducing this class of models in our work. In many practical bandit scenarios, adoption of classical bandit algorithms designed for a stationary environment seems commonplace, despite the fact that the environment could be nonstationary as a result of being dependent on the population wide actions. This observation begs a couple of natural questions: Does it matter that agents make the stationarity assumption? What are the conditions under which the environment does become stationary? We make significant progress in providing these answers by studying the MAB game in a *mean field* regime, inspired by an approximation where the number of agents becomes large. In particular, suppose agents' conjecture that competitors pull arms at a frequency given by their long run average; we refer to this long run average of arm frequencies across agents as the *population profile*. Under this conjecture, their environment appears stationary, so the agent's optimization problem becomes a classical MAB problem. Of course, a consistency check is required: the conjectured population profile must arise from agents' chosen policy. We refer to the resulting fixed point as a *mean field steady state* (MFSS).

Our main results are as follows.

1. *Existence of MFSS for MAB games.* In the model we consider, we assume that agents play a fixed policy; for example, this may be a regret-optimal policy for the classical (stationary) MAB setting. We establish existence of MFSS for this model. While fixing the policy agents use may be ill justified for an equilibrium concept like the PBE, this approach is sensible for MAB games for reasons outlined above; for example, if agents use a regret-minimizing policy (such as UCB), we can show that it is approximately optimal for an MAB game.

2. *Uniqueness and convergence.* We identify a *contraction* condition on the arm rewards that ensures the MFSS is unique, and that starting from any initial state, the dynamics will converge to

this MFSS (in the sense that eventually the population profile becomes constant). The contraction condition requires that the agent population is sufficiently mixing and that the sensitivity of the reward function to the population profile is low enough.

We use numerical experiments to investigate the tightness of this condition. We find that in settings with *positive externalities* among the agents, our condition is relatively tighter: violations of the condition lead to nonconvergence of dynamics. On the other hand, with *negative externalities*, the system converges even well outside the regime specified by our condition.

3. *Approximation.* Under the same contraction condition used to establish uniqueness, we show an approximation result that justifies our use of MFSS. In particular, we establish that if the number of agents grows large, then the dynamics of the finite agent system converge to the dynamics of the mean field model.

The remainder of the paper is organized as follows. In Section 2, we discuss related literature. In Section 3, we introduce our mean field model, and basic notation and definitions. In Section 4, we consider two classes of applications that can be captured by our model: those that exhibit *negative* externalities, and those that exhibit *positive* externalities. In Section 5, we introduce the notion of mean field steady state, and establish existence. In Section 6, we show under the contraction condition that the MFSS is unique, and that dynamics converge to it starting from any initial condition. In Section 7, we establish an approximation theorem that justifies MFSS; in particular, we show that when the number of agents grows large, the behavior of the finite system approaches the behavior of the mean field model. In Section 8, we present numerical experiments used to study MFSS when the contraction condition is violated. Proofs of the main theorems are provided in the appendix. The appendix also contains two extensions of our main results. First, we show that the same results hold even when agents may use heterogeneous policies. Second, we show that when negative externalities are present, and agents use policies that exhibit positive sensitivity to arm rewards (in a sense we make explicit), then MFSS are unique.

2. Related work Our work is related to two different strands of research: the first on multi-armed bandits, and the second on mean field equilibria of dynamic games. We briefly survey each line of work here.

Multi-armed bandits have a long history, having first been introduced in [35] and arise frequently in many settings, ranging from online ad auctions to clinical trials. They have been the focus of attention from many different research communities, including control theory, statistics, computer science, and operations research. Several different variants have been considered; for example, the stochastic optimal control setting, e.g., [20, 41, 28, 38, 33] and the references therein; the regret performance criterion in the stochastic model, e.g., [29, 5, 6, 3, 8, 25] and references therein; and the adversarial model, including the weighted majority algorithm and generalizations, e.g., [32, 10, 7] and the references therein. A recent line of work on bandit algorithms has highlighted the importance of a class of algorithms based on posterior sampling [37, 13, 4, 34, 36, 21]. Our work is complementary to this line of work: the multi-armed bandit problem is a building block of the game theoretic model we study, and most of our results are agnostic to the specific policy applied.

The second line of work we briefly review concerns the use of mean field models in dynamic games. This work dates back to [26] and [22], but has seen a recent surge of theoretical development; see, e.g., [30, 23, 2]. In addition to this literature, mean field equilibria have been successfully studied in a range of applications, including dynamic auctions [24], dynamic oligopoly models [39, 40], dynamic search [16], coupling of oscillators [43], large markets [11], transportation systems [17], games with complementarities [1] and on particle filters [42]. Our work complements this literature by applying mean field equilibria to obtain insight into multi-armed bandit games.

3. Model We begin by considering a model where the number of agents is finite; the "mean field" model can be viewed as a formal specification of this model in the regime where the agent population is infinite. We refer to the model presented in this section as the *finite agent system*. We consider a setting where each agent is solving a multi-armed bandit (MAB) problem, but where the rewards of arms depend on the actions of other agents.

Time. Time is discrete and indexed by $t \in \{0, 1, \ldots\}$.

Agents. We assume that there are a total of m agents in the system, where the agents are indexed by $k \in \{1, \ldots, m\}$. Throughout the paper we typically use the superscript m to denote quantities in the finite agent system with m agents, as we later develop asymptotic analysis of the system as $m \to \infty$; when it is clear from context, we suppress the superscript for notational simplicity.

Arms. The number of arms is n; these correspond to the actions an agent can choose at any given time. We assume that the outcome when an agent pulls an arm is binary: a "win" corresponds to a unit reward received from the pull of the arm, and a "loss" corresponds to a zero reward received from the pull of the arm. Thus the relevant statistic for the agent will be the probability that pulling a given arm yields a win.

Types. Each agent k has a *type*, given by a random variable $\boldsymbol{\theta}(k)$ taking values in $[0, 1]^n$. The n components of $\boldsymbol{\theta}(k)$ represent parameters that influence the reward distributions across the n arms. Thus $\boldsymbol{\theta}(k)$ captures *agent-specific* variation in the relative values of different arms. Note that the type of the agent is unobservable and is among the parameters the agent must learn through exploration.

We assume that across the agents, $\boldsymbol{\theta}(k)$ is independent and identically distributed. The population measure of $\boldsymbol{\theta}(k)$, which is unknown to the agents, is denoted by W, where $W(\Theta)$ represents the probability that $\boldsymbol{\theta}(k) \in \Theta \subseteq [0, 1]^n$. The evolution of $\boldsymbol{\theta}(k)$ across time for each agent is described in the regeneration model for agents below.

States. In principle, an agent can use a policy that depends on the entire *history* of past play (i.e., the sequence of wins and losses during their lifetime). Unfortunately, this state description grows exponentially with time, and thus it seems unreasonable to consider strategies that act at this level of detailed historical information.

For our purposes, therefore, we restrict attention to a simpler model of the agent. In particular, we define the *state* of agent k at time t to be given by the vector $\boldsymbol{z}_t(k) \in \mathbb{Z}_+^{2n}$, where the $2n$ elements correspond to the *total* wins and losses on each of the n arms, over the lifetime of the agent. Thus we lose the information about *when* wins and losses were obtained in the past. Many policies of interest act on this level of aggregated state information, and thus it is a reasonable assumption for our setting. We note in contrast to the type, the state $\boldsymbol{z}_t(k)$ *is* observed by the agent at each time and thereby directly influences the action.

Policies. As suggested above, the agents are presumed to choose an arm to play at each time according to a stationary policy that only depends on the current state. We assume all agents use the same policy; while this assumption may seem restrictive, it can be relaxed by considering a model where the type also consists of an observable component that encodes the strategy a particular agent chooses to use. This extension is outlined in more detail in Section B.

We let $S = \{x \in [0, 1]^n : \sum_{i=1}^n x_i = 1\}$; this is the set of randomized actions over the arms. The policy used by the agents is denoted by $\sigma : \mathbb{Z}_+^{2n} \to S$. For notational convenience, for any $\boldsymbol{z}$, we let $\sigma(\boldsymbol{z}, i)$ denote the probability of an agent choosing arm i while in state $\boldsymbol{z}$. Therefore, we have $\sum_{i=1}^n \sigma(\boldsymbol{z}, i) = 1$ for all states $\boldsymbol{z}$. We denote the specific realization of the arm choice of agent k at time t while in state $\boldsymbol{z}$ by the random variable $\sigma_t(k, \boldsymbol{z}) \in \{1, \ldots, n\}$. For any fixed $\boldsymbol{z}$, $\sigma_t(k, \boldsymbol{z})$ is i.i.d. over all pairs (k, t), with its probability distribution specified by $\mathbb{P}(\sigma_t(k, \boldsymbol{z}) = i) = \sigma(\boldsymbol{z}, i)$ for all k, $\boldsymbol{z}$, and i.

As noted above, this model is fairly general from the perspective of the policy being adopted. For example, suppose that the agent's goal is to maximize the expected reward over their lifetime

with respect to some initial prior. Because of the regeneration at geometric rate, this expected reward maximization is equivalent to maximization of expected discounted reward over an infinite horizon, for the which the corresponding optimal Gittins index policy is a function of the state vector [20, 41, 28, 38]. More generally, the posterior density at each given time is a function of the state vector; and therefore, any Bayesian methods that depend on the posterior (e.g. [37, 13, 4, 34, 36, 21]) depend only on the current state in our model. Other policies that also fall into this class include those based on the principle of optimism in the face of uncertainty, introduced by [29] and its generalizations to algorithms based on Upper Confidence Bounds (UCB) and its variants [3, 12, 8, 9]. From now on, we abstract away the specific policy used by representing it with the map σ as defined.

Regeneration. A key assumption we make is that each agent is *regenerated* independently with probability $1-\beta$ in each time slot; thus an agent has a geometric lifetime before regeneration. At regeneration of agent k, $\boldsymbol{z}_t(k)$ is re-initialized to the all-zero vector, and $\boldsymbol{\theta}_t(k)$ is sampled (independently of all other randomness) according to the distribution W. Between regenerations, an agent's type remains fixed.

We assume regeneration for two reasons. First, regeneration naturally models a process of arrival and departure of agents to the system, and make sense for systems with dynamic populations. Second, from an analytical standpoint, if our agent population never regenerated, then we should expect that in steady state no learning is necessary: under reasonable conditions, agents will resolve all type uncertainty. As a result, in steady state the model would reduce to a game without any learning required on the part of the agents. By including regeneration, we ensure that even in steady state, agents must continue to learn. (A similar assumption is made by [24] in their analysis of dynamic auctions.)

Occupation measure and population profile. We refer to the population-wide distribution over states and types as the *occupation measure*. Before introducing the definition, we require one additional piece of notation. Let $\mathcal{M}$ denote the space of all (Borel) probability measures on $\mathbb{Z}_+^{2n} \times [0,1]^n$. Thus $\mathcal{M}$ describes the set of all possible joint probability distributions over agent states and types. For notational convenience (since states are discrete and types are continuous), for any $\mu \in \mathcal{M}$, we write $\mu(\boldsymbol{z}, \Theta)$ for the probability of state $\boldsymbol{z}$ cross with the type set Θ.

The (random) occupation measure of the finite agent system with m agents at time t is denoted by $\mu_t^m \in \mathcal{M}$; for a state $\boldsymbol{z}$ and $\Theta \subseteq [0,1]^n$, we have:

$$\mu_t^m(\boldsymbol{z}, \Theta) \triangleq \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\boldsymbol{z}_t^m(k) = \boldsymbol{z}, \boldsymbol{\theta}_t^m(k) \in \Theta\}.$$

The *population profile* is defined to be the histogram of the arm choices realized by the agent population. The (random) empirical population profile at time t in the m'th system is denoted by $\boldsymbol{f}_t^m$, and is specified based on the arm choice realization of each agent, together with the occupation measure. In particular, for each i, we have:

$$f_t^m(i) \triangleq \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\sigma_t(k, \boldsymbol{z}_t^m(k)) = i\}.$$

Rewards. We assume that rewards are conditionally independent (across arms, time, and agents), given an agent's type and the population profile component of the arm pulled by that agent. In particular, we assume that an agent receives a unit reward (a "win") with probability $Q(\theta(i), f(i))$ when she pulls arm i, if her type is $\boldsymbol{\theta}$ and the population profile is $\boldsymbol{f}$. Throughout the paper, we assume that *for each fixed θ, $Q(\theta, \cdot)$ is continuous.*

State and type transitions. Depending on whether the reward variable corresponding to the chosen arm is one or zero, the player's state corresponding to the appropriate variable increases

by one, conditional on not regenerating in the next time slot. More precisely, let $\boldsymbol{w}_i$ and $\boldsymbol{l}_i$ denote unit basis vectors corresponding to a win or loss, respectively, on pulling arm i. Thus if the current state of an agent is $\boldsymbol{z}(k)$, the state updates to $\boldsymbol{z}+\boldsymbol{w}_i$ if arm i is pulled and wins, or $\boldsymbol{z}+\boldsymbol{l}_i$ if arm i is pulled and loses.

For an agent using a fixed policy σ, the state-type vector transitions following the kernel K defined below. Suppose the agent's current type is $\boldsymbol{\theta}$; the state is $\boldsymbol{z}$; and the population profile is $\boldsymbol{f}$. Then:

$$\begin{aligned}\mathbb{K}\left(\boldsymbol{z}',\Theta\middle|\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{f},\sigma\right) &\triangleq \mathbb{P}\left(\boldsymbol{z}_{t+1}=\boldsymbol{z}',\boldsymbol{\theta}_{t+1}\in\Theta\middle|\boldsymbol{z}_t=\boldsymbol{z},\boldsymbol{\theta}_t=\boldsymbol{\theta};\boldsymbol{f},\sigma\right)\\ &=(1-\beta)\mathbb{1}_{\boldsymbol{z}'=0}W(\Theta)\\ &\quad+\beta\mathbb{1}_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n}\sigma(\boldsymbol{z},i)\left(Q(\theta(i),f(i))\mathbb{1}_{\boldsymbol{z}'=\boldsymbol{z}+\boldsymbol{w}_i}+(1-Q(\theta(i),f(i)))\mathbb{1}_{\boldsymbol{z}'=\boldsymbol{z}+\boldsymbol{l}_i}\right).\end{aligned}\tag{1}$$

In the transition kernel, the first term corresponds to regeneration, where with probability $1-\beta$, the state resets to zero, and the type is independently resampled from the distribution $W(\Theta)$. The second term represents the case where regeneration does not occur, so the state evolves in response to a choice of one of the n possible arms. In the current state $\boldsymbol{z}$, arm i is chosen according to the policy σ, with probability $\sigma(\boldsymbol{z},i)$. The resulting pull is successful with probability $Q(\theta(i),f(i))$, and the state vector updates in response to the resulting win or loss.

System dynamics. Finally, using the preceding definitions, we can describe the stochastic dynamics of the system as a whole. The initial state of each agent at time 0 is set to $\boldsymbol{z}=0$, and the type of each agent at time 0 is sampled independently from W. Given $\boldsymbol{f}_t^m$, an agent of type $\boldsymbol{\theta}$ at state $\boldsymbol{z}$ transitions independently following the kernel $\mathbb{K}\left(\cdot\middle|\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{f}_t^m,\sigma\right)$, where $\mathbb{K}$ is defined in Equation 1. With these definitions, the occupation measure of the m'th system μ_t^m is a Markov process with state space $\mathcal{M}$.

4. Examples Central to our model is the assumption that arm rewards depend on both an agent specific parameter, as well as the population profile of other agents' decisions. In this section, we motivate our model by considering two broad classes of reward structure: (1) *negative externalities*, where agents' rewards are decreasing as the fraction of other agents choosing the same arm increases; and (2) *positive externalities*, where agents' rewards are increasing as the fraction of other agents choosing the same arm increases. Below we provide both examples of applications where each type of structure might arise, as well as specific functional forms for the rewards that exhibit each type of behavior.

4.1. Negative externalities Negative externalities arise when agents suffer lower rewards due to the actions of other agents. In our context, this might be modeled by assuming that $Q(\theta,f)$ is decreasing in f. Here we discuss a few examples where such behavior arises, as well as functional forms that exhibit this behavior.

First, consider a wireless network with n distinct frequency bands. Suppose the agents are individual devices attempting to transmit a packet in each time slot by accessing the spectrum on one of these bands, or channels. In this case, θ corresponds to agent specific channel quality (i.e., channel gain), and f influences the interference created by other agents using the same band. Thus a reasonable model is one where $Q(\theta,f)$ is the channel success probability for a given agent, and thus is decreasing in f. Further, regeneration in this context might correspond naturally to changes in the agent specific channel quality over time.

As a second example, consider a competitive scenario where firms or agents compete for success in n different product markets; one example might be an online marketplace such as eBay, where

sellers can compete in any of n different categories (or perhaps on any of several other dimensions of product characteristics). In this case, θ might represent the agent specific appropriateness, while f represents the competition from other sellers in the same category. Thus, again we might expect $Q(\theta, f)$ to be decreasing in f. This example also illustrates the ways in which our model is stylized; for example, in a model of eBay, reputation feedback effects will also affect seller behavior. Nevertheless, we feel the general class of model studied here can serve as a baseline from which to explore more complex application-dependent models.

As noted, the models above all point to specific forms of Q that are decreasing in the population profile. As one simple functional form, consider the case where $Q(\theta, f) = \theta G(f)$ for some decreasing function $G(f)$ (taking values in $[0, 1]$). In our numerical examples, in particular, we will consider such forms where $G(f) = 1/(1 + Lf)$ where $L \geq 0$.

4.2. Positive externalities Positive externalities arise in scenarios where agents' rewards increase with the number of other agents that choose the same action. Such models often arise in the context of cooperation or coordination games, or as we discuss here, when there are strong *network effects*.

Online social gaming scenarios provide one example of such settings. Suppose that an individual has the choice of n different games they may participate in. Agents' types in this setting corresponds to their intrinsic preference for a given game; but their rewards are tied not only to their own participation in the game, but also to how many other agents choose the same game to play as well. In this case an agent's reward will be increasing in the fraction of other agents who choose the same game.

One possible functional form, in line with the kinds of examples discussed above, is again $Q(\theta, f) = \theta G(f)$, but where now $G(f)$ is an increasing function (taking values in $[0, 1]$). For example, we might simply consider $G(f) = f$. On the other hand, such a form has the flaw that it does not allow the agent to derive any benefit if f is zero. An alternative reward structure might be $Q(\theta, f) = F(\theta) + \theta G(f)$, which allows the possibility for agent-specific rewards even in the absence of participation from other agents.

5. Mean field model and mean field steady state In this section, we introduce a formal *mean field* model, meant to capture the dynamics of the system in a setting where the number of agents grows large. In this section and the next, we treat this as a formal construct. In Section 7, we provide conditions under which the interpretation of the mean field model as a limit of the finite agent system is justified.

The mean field model consists of a sequence of occupation measures $\mu_0, \mu_1, \ldots$, with associated population profile sequence $\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots$. (Note that these objects do not have a superscript m, to emphasize that they are part of the mean field model.) Informally, the idea is the following: the measure of agents across the population is given by μ_t at time t. We imagine each of these agents playing according to the fixed policy σ, and earning rewards given the population profile of actions $\boldsymbol{f}_t$ at that time; each agent then transitions according to the kernel $\mathbb{K}(\cdot|\boldsymbol{\theta}, \boldsymbol{z}; \boldsymbol{f}_t, \sigma)$.

In the formal mean field model, we do not explicitly model the state dynamics of every individual agent. Rather, we capture the preceding discussion formally by defining a dynamical system $\{\mu_t\}_{t\geq 0}$. This system is a sequence of deterministically evolving measures over time, together with an associated sequence of population profiles $\{\boldsymbol{f}_t\}_{t\geq 0}$. We assume the initial condition is that agents start with the regeneration measure at time 0, i.e., $\mu_0(\boldsymbol{z}, \Theta) \triangleq \mathbb{1}_{\boldsymbol{z}=\boldsymbol{0}} W(\Theta)$. Next, the sequence of measures μ_t are defined iteratively as $\mu_{t+1} = \mathcal{T}_\sigma(\mu_t)$, where:

$$\mathcal{T}_\sigma(\mu)(\boldsymbol{z}, \Theta) \triangleq \int_{(\boldsymbol{z}', \boldsymbol{\theta})} \mathbb{K}\left(\boldsymbol{z}, \Theta \middle| \boldsymbol{z}', \boldsymbol{\theta}; p(\sigma, \mu_t), \sigma\right) d\mu, \tag{2}$$

where $\mathbb{K}$ is defined in Equation 1, and where:

$$p(\sigma,\mu)(i) \triangleq \sum_{\boldsymbol{z}} \sigma(\boldsymbol{z},i)\mu(\boldsymbol{z},[0,1]^n). \tag{3}$$

We can interpret p as the marginal distribution over arms when the joint state and type distribution is μ, and the policy used is σ; in particular, given the occupation measure μ_t at time t, the corresponding population profile is defined as:

$$\boldsymbol{f}_t = p(\sigma,\mu_t).$$

Each agent then transitions with this population profile driving their state transitions, and assuming they follow the policy σ. The map $\mathcal{T}_\sigma$ then integrates these state transitions over the entire occupation measure μ to obtain the next iterate in the dynamical system.

We refer to fixed points of the map $\mathcal{T}_\sigma$ as *mean field steady states* (MFSS). Informally, these are "equilibria" of the system. Given the preceding discussion, it is useful to think of an MFSS as encoding two complementary conditions that we refer to as *stationarity* and *consistency*, as in the following definition.

DEFINITION 1. Given a type distribution W, a reward function Q, and a policy σ, a *mean field steady state* is a pair $(\mu,\boldsymbol{f})$ where $\mu \in \mathcal{M}$ is a joint distribution on states and types, and $\boldsymbol{f}$ is a population profile, such that the following two conditions hold:

C1 *Stationarity*: For any state $\boldsymbol{z}$ and any type set Θ, there holds:

$$\mu(\boldsymbol{z},\Theta) = \int_{\boldsymbol{z}',\boldsymbol{\theta}} \mathbb{K}\left(\boldsymbol{z},\Theta \middle| \boldsymbol{z}',\theta;\boldsymbol{f},\sigma\right) d\mu;$$

C2 *Consistency*: There holds $\boldsymbol{f} = p(\sigma,\mu)$, where p is defined as in (3).

The first condition requires that μ is the steady state distribution of the agent state dynamics induced when the population profile is $\boldsymbol{f}$, and all agents use policy σ. The second condition requires that $\boldsymbol{f}$ is exactly the arm distribution induced when the steady state distribution of the population is μ, and all agents use policy σ. The two conditions together ensure that the system is in equilibrium.

Note that if μ is a fixed point for the map $\mathcal{T}_\sigma$, then μ and $\boldsymbol{f} = p(\sigma,\mu)$ are an MFSS. Conversely, in any MFSS $(\mu,\boldsymbol{f})$, μ must be a fixed point of $\mathcal{T}_\sigma$ (and $\boldsymbol{f}$ is uniquely determined as $\boldsymbol{f} = p(\sigma,\mu)$).

We have the following basic proposition, that establishes existence of MFSS. The proof is via a topological fixed point argument, and relies on our assumption that Q is continuous in the population profile. We defer the proof to the appendix.

PROPOSITION 1. *For any policy σ, there exists a MFSS $(\mu,\boldsymbol{f})$.*

Recall that our motivation behind considering this concept is to understand whether policies that are optimal for a stationary MAB problem lead to stationarity even in dynamic environments that allow for time-varying population profiles. As such, we refrain from make any specific additional assumptions regarding the optimality of the policy. In our approach, we allow a general family of policies, and hold the policy fixed when proving existence of an MFSS. The definition can be refined by asking that agents use optimal or approximately optimal policies; that is, we might require that in addition to C1 and C2 above, agents' use a policy that maximizes (or approximately maximizes) their expected reward over their lifetime in the system. Observe that any such MFSS would also be an MFSS in our definition, with a particular choice of policy σ. This refinement is not needed for our development below, as we derive results that are agnostic to the particular policy chosen.

We also emphasize that an important feature of the model is that agents are oblivious to the parameters. In other words, an agent can only learn about its environment through the history

of rewards accumulated during the current regeneration lifetime. An optimization of the objective function over the space of policies can therefore lend itself to different interpretations, depending on how we model the agents. For example, one possibility is to endow the agents with uniform beta priors for the arm reward distributions. In this case, an index policy will be optimal for the agents (as previously discussed). Alternately, agents might choose a regret minimization algorithm over a geometric random horizon as an approximately optimal policy. One virtue of our notion of equilibrium is that it is flexible with respect to the specific assumptions behind the agents' notion of optimality.

We note here one example of how policies might be chosen that are approximately optimal. In particular, suppose that agents use a *no-regret policy* for the usual MAB problem with stationary reward distributions. For a fixed time T, let $\boldsymbol{z}_t : t \in \{0, 1, \ldots, T-1\}$ denote the states visited by a given agent under the policy σ, *when the population profile is stationary at* $\boldsymbol{f}$. Suppose that for every T, the policy σ satisfies:

$$\max_i Q(\theta(i), f(i)) - \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} Q(\theta(\sigma(\boldsymbol{z}_t)), f(\sigma(\boldsymbol{z}_t)))\right] < \frac{R(T)}{T}, \tag{4}$$

for some function R; $R(T)/T$ gives an upper bound to the average regret under σ.

Now let T be the regeneration lifetime of a given agent; specifically, T is geometric with parameter β. In this case the policy σ is ϵ-*optimal*, in the sense that:

$$\max_i Q(\theta(i), f(i)) - \mathbb{E}_T\left[\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n} Q(\theta(i), f(i))\sigma(\boldsymbol{z}_t, i)\right] < \epsilon,$$

for the particular choice of ϵ given by:

$$\epsilon = \sum_{T=1}^{\infty}(1-\beta)\beta^{T-1}\frac{R(T)}{T}.$$

We assume here that the sum on the right is finite; the result follows by taking an expectation over T in (4). For the algorithms based on Upper Confidence Bounds (UCB) [9] or Thompson sampling [21] for example, we have logarithmic regret bounds under a very broad class of models: $R(T) = \alpha \ln T$ for some constant α, which is also optimal with a lower bound of matching order [29]. For $\beta \to 1$, ϵ scales like $-\alpha(1-\beta)\log(1-\beta)$. This example reveals that when using no-regret policies, any MFSS has the property that agents play approximately optimally.

6. Mean field steady state: Uniqueness and convergence In this section we consider conditions under which MFSS is unique. Uniqueness of equilibria is useful primarily because it lends greater predictive value to the steady state concept. In our case, however, uniqueness also yields significant additional insight: the same conditions we use to establish uniqueness of MFSS are used to establish convergence of out-of-steady-state dynamics to MFSS. Furthermore, in the next section, these conditions are also used to establish that the behavior of systems with a finite number of agents approaches the mean field model as the number of agents grows.

Consider the endowment of $\mathcal{M}$ with the total variation distance metric, d_{TV}. We make use of the following equivalent characterizations of d_{TV}.

Definition 2 ([19, 31]). The total variation distance (which metrizes $\mathcal{M}$) has the following equivalent definitions:

D1 Let $\mu_1, \mu_2 \in \mathcal{M}$ be any two measures and let μ be any measure which is absolutely continuous with respect to both μ_1 and μ_2. Let μ_1', μ_2' be their Radon-Nikodym derivatives with respect to μ. Then:

$$d_{TV}(\mu_1, \mu_2) \triangleq \frac{1}{2}\int_{(z,\theta)} |\mu_1'(z,\theta) - \mu_2'(z,\theta)| d\mu.$$

D2 Let F denote the set of all (Borel) measurable subsets of $\mathbb{Z}_+^{2n} \times [0,1]^n$. Then:

$$d_{TV}(\mu_1, \mu_2) = \sup_{A \in F} |\mu_1(A) - \mu_2(A)|.$$

D3 Let $\Omega(\mu_1, \mu_2)$ denote all joint probability measures on two random vectors $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$ such that the marginal distribution of $(\boldsymbol{z}_k, \boldsymbol{\theta}_k)$ is μ_k for $k = 1, 2$. Then:

$$d_{TV}(\mu_1, \mu_2) = \inf_{\Omega(\mu_1, \mu_2)} \mathbb{P}\big((\boldsymbol{z}_1, \boldsymbol{\theta}_1) \neq (\boldsymbol{z}_2, \boldsymbol{\theta}_2)\big).$$

Given any policy σ, we derive a sufficient condition for the existence of a *unique* MFSS (μ, f), and for convergence to that MFSS. Recall that MFSS are equivalent to fixed points of the map $\mathcal{T}_\sigma$, cf. (2). Therefore, given a policy σ, a unique MFSS exists if and only if the map $\mathcal{T}_\sigma$ has a unique fixed point. A sufficient condition for this is that $\mathcal{T}_\sigma$ is a *contraction* map with respect to the total variation distance metric, d_{TV}. This condition also implies that if we initialize the mean field dynamical system from an arbitrary initial distribution $\mu \in \mathcal{M}$, then eventually the resulting dynamics converge to an MFSS.

The next theorem provides one sufficient condition for $\mathcal{T}_\sigma$ to be a contraction map with respect to d_{TV}. This in turn, establishes that the MFSS is unique, and that the mean field dynamics converge to it. We make two remarks on the result. First, observe that the dynamics described here are easy to interpret. The system begins in some out-of-equilibrium joint distribution over agents' types and states. Each agent applies σ each period, to the history they have seen since regeneration. As the system evolves forward, the population profiles will change at each time t, and this in turn affects the reward sequence seen by an agent. However, the theorem shows that despite the interaction across agents, the environment does become stationary. More precisely, the joint distribution over agents' types and states will converge to μ^*, and the population profile will converge to the corresponding f^*—the unique MFSS. In this way our result demonstrates conditions under which even when agents apply policies derived in a single agent environment, the resulting multiagent dynamics (in the mean field limit) become stable.

Our second remark concerns the bound $\beta(1+L) < 1$ required in our theorem. This bound suggests a tradeoff: the rate at which agents regenerate should not be too high, and the arm rewards should not vary too much with respect to changes in the population profile. These conditions point to scenarios where agents are coupled, but the coupling is not intense and does not last too many periods. Indeed, the condition is in some sense quite restrictive, since very high β (corresponding to longer lifetimes) requires Q to be relatively insensitive to the population profile. We investigate how restrictive the condition is in practice through numerical experiments in Section 8.

THEOREM 1. *Suppose that for all $a \in [0,1]$ and $x, x' \in [0,1]$, there holds:*

$$|Q(a,x) - Q(a,x')| \leq L|x - x'|. \tag{5}$$

Then for any policy σ, $\mathcal{T}_\sigma$ is a contraction map with respect to the total variation distance, d_{TV} when the following condition holds:

$$\beta(1+L) < 1. \tag{6}$$

In particular, there exists a unique fixed point μ^* of T_σ, and thus a unique MFSS. Further, starting from any initial distribution $\mu_0 = \mu$, if agents all use policy σ, the mean field dynamical system $\mu_0, \mu_1, \cdots$ defined by $\mu_{t+1} = \mathcal{T}_\sigma(\mu_t)$ converges to to μ^* (in total variation distance).

Proof. Fix $\mu_1, \mu_2 \in \mathcal{M}$ such that $d_{TV}(\mu_1, \mu_2) = d$. We prove the theorem by showing that:

$$d_{TV}(\mathcal{T}_\sigma(\mu_1), \mathcal{T}_\sigma(\mu_2)) \leq \beta d(1+L).$$

From characterization D3 of Proposition 2, our desired inequality would follow if we construct two random variables, $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$ on a common probability space such that:

1. $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ has distribution $\mathcal{T}_\sigma(\mu_1)$;
2. $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$ has distribution $\mathcal{T}_\sigma(\mu_2)$; and
3. $\mathbb{P}\big((\boldsymbol{z}_1, \boldsymbol{\theta}_1) \neq (\boldsymbol{z}_2, \boldsymbol{\theta}_2)\big) \leq \beta(1+L)d$.

Since $d_{TV}(\mu_1, \mu_2) = d$, again from characterization D3 of Proposition 2, fix $\delta > 0$ and let $(\boldsymbol{z}'_1, \boldsymbol{\theta}'_1)$ and $(\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)$ be two random variables defined on a common probability space such that:

1. $(\boldsymbol{z}'_1, \boldsymbol{\theta}'_1)$ has distribution μ_1;
2. $(\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)$ has distribution μ_2; and
3. $\mathbb{P}\big((\boldsymbol{z}'_1, \boldsymbol{\theta}'_1) \neq (\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)\big) = d + \delta$.

To construct the desired coupled random variables, we define coupled *transitions* from $(\boldsymbol{z}'_1, \boldsymbol{\theta}'_1)$ to $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)$ to $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$. In particular:

1. Let $\hat{\beta}$ be an independent Bernoulli random variable equal to one with probability β, and zero otherwise.
2. Let $\hat{\boldsymbol{\theta}} \in [0,1]^n$ be a independent random variable sampled from the distribution W.
3. Let $\hat{\sigma}(\boldsymbol{z}) \in \{1, \ldots, n\}$ be independent random variables for each $\boldsymbol{z} \in \mathbb{Z}_+^{2n}$ such that $\mathbb{P}(\hat{\sigma}(\boldsymbol{z}) = i) = \sigma(\boldsymbol{z}, i)$ for all i.

To economize on notation below, we write $\hat{a}_k = \hat{\sigma}(\boldsymbol{z}'_k)$ for $k = 1, 2$. Observe with this definition that if $\boldsymbol{z}'_1 = \boldsymbol{z}'_2$, then $\hat{a}_1 = \hat{a}_2$.

4. Let $\hat{u} \in [0,1]$ be a uniform random variable independent of all other randomness. For $k = 1, 2$, define:

$$\hat{Q}_k = \begin{cases} 1 & \text{if } u \leq Q(\theta'_k(\hat{a}_k), p(\sigma, \mu_k)(\hat{a}_k)); \\ 0 & \text{otherwise.} \end{cases}$$

Note that the single choice of u couples together transitions in both systems.

Now we couple $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$ as follows. For $k = 1, 2$ define:

$$\begin{aligned} \boldsymbol{\theta}_k &= \hat{\beta}\boldsymbol{\theta}'_k + (1 - \hat{\beta})\hat{\boldsymbol{\theta}}; \\ \boldsymbol{z}_k &= \hat{\beta}(\boldsymbol{z}'_k + \hat{Q}_k \boldsymbol{w}_{\hat{a}_k} + (1 - \hat{Q}_k)\boldsymbol{l}_{\hat{a}_k}) + (1 - \hat{\beta})\mathbf{0}. \end{aligned}$$

This coupling has the following features: (i) regeneration is common in both systems; (ii) and informally, the evolution of the state is as common "as possible", in the sense that state transitions are driven by common randomness. It can now be verified by our construction that $(\boldsymbol{z}_1, \boldsymbol{\theta}_1)$ and $(\boldsymbol{z}_2, \boldsymbol{\theta}_2)$ have the distributions $\mathcal{T}_\sigma(\mu_1)$ and $\mathcal{T}_\sigma(\mu_2)$ respectively.

We start by noting that by construction, $\mathbb{P}(\hat{a}_1 \neq \hat{a}_2) \leq \mathbb{P}(\boldsymbol{z}'_1 \neq \boldsymbol{z}'_2) \leq d + \delta$. Since $p(\sigma, \mu_1), p(\sigma, \mu_2)$ are the respective probability distributions of $\hat{a}_1$ and $\hat{a}_2$, using characterization D3 again, we have $d_{TV}(p(\sigma, \mu_1), p(\sigma, \mu_2)) \leq d$. Using characterization D2, we conclude that for all i,

$$|p(\sigma, \mu_1)(i) - p(\sigma, \mu_2)(i)| \leq d + \delta. \tag{7}$$

We proceed to upper bound the probability of the event $E \triangleq \{(\boldsymbol{z}_1, \boldsymbol{\theta}_1) \neq (\boldsymbol{z}_2, \boldsymbol{\theta}_2)\}$. Let E' be defined by $E' \triangleq \{(\boldsymbol{z}'_1, \boldsymbol{\theta}'_1) = (\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)\}$. On E', let $(\boldsymbol{z}', \boldsymbol{\theta}') \triangleq (\boldsymbol{z}'_1, \boldsymbol{\theta}'_1) = (\boldsymbol{z}'_2, \boldsymbol{\theta}'_2)$, and let $\hat{a} \triangleq \hat{a}_1 = \hat{a}_2$. Then we have from from the assumption on Q with $\zeta_1 = p(\sigma, \mu_1)(\hat{a}), \zeta_2 = p(\sigma, \mu_2)(\hat{a})$, that:

$$|Q(\theta'(\hat{a}), \zeta_1) - Q(\theta'(\hat{a}), \zeta_2)| \leq L\,|p(\sigma, \mu_1)(\hat{a}) - p(\sigma, \mu_2)(\hat{a})|\,.$$

Using the inequality (7) with $i = \hat{a}$, and combining it with the preceding Lipschitz inequality, the coupling between $\hat{Q}_1$ and $\hat{Q}_2$ implies:

$$\begin{aligned} \mathbb{P}\left(\hat{Q}_1 \neq \hat{Q}_2 \middle| E', \boldsymbol{z}'\right) &= |Q(\theta'(\hat{a}), p(\sigma, \mu_1)(\hat{a})) - Q(\theta'(\hat{a}), p(\sigma, \mu_2)(\hat{a}))| && (8) \\ &\leq L\,|p(\sigma, \mu_1)(\hat{a}) - p(\sigma, \mu_2)(\hat{a})| \leq L(d + \delta). && (9) \end{aligned}$$

It follows that $\mathbb{P}(\hat{Q}_1 \neq \hat{Q}_2 | E') \leq L(d + \delta)$.

To conclude the proof, note from our coupling that:

$$E \subseteq \{\hat{\beta}=1\} \bigcap \left\{ E'^{C} \bigcup \left\{ E' \bigcap \left\{ \hat{Q}_1 \neq \hat{Q}_2 \right\} \right\} \right\}.$$

In other words: if $(\boldsymbol{z}_1, \boldsymbol{\theta}_1) \neq (\boldsymbol{z}_2, \boldsymbol{\theta}_2)$, there must not be a regeneration (since the coupling ensures post-regeneration type and state are identical in both systems); and either the initial states and/or types differ, or after the transition in each system, the subsequent states differ. Bounding the probability of the RHS event using the bound obtained in the preceding paragraph, we obtain:

$$\begin{aligned} \mathbb{P}(E) &\leq \beta \left(\mathbb{P}(E'^{c} | \hat{\beta}=1) + \mathbb{P}(E' | \hat{\beta}=1) \cdot \mathbb{P}(\hat{Q}_1 \neq \hat{Q}_2 | E', \hat{\beta}=1) \right) && (10) \\ &\leq \beta(d+\delta+L(d+\delta)) = \beta(1+L)(d+\delta). && (11) \end{aligned}$$

THe second line follows because the random variable $\hat{\beta}$ (determining regeneration) is independent of the event E' (which depends only on the initial states in the two systems). The preceding derivation implies that for any arbitrary μ_1, μ_2, we have:

$$d_{TV}(\mathcal{T}_\sigma(\mu_1), \mathcal{T}_\sigma(\mu_2)) \leq \beta(1+L)(d+\delta).$$

Since δ was arbitrary, we can take $\delta \to 0$ and conclude that:

$$d_{TV}(\mathcal{T}_\sigma(\mu_1), \mathcal{T}_\sigma(\mu_2)) \leq \beta(1+L)d,$$

as required.

This implies that under the given assumption, $\mathcal{T}_\sigma$ is a contraction map on the space of probability measures on $\mathbb{Z}_+^{2n} \times [0,1]^n$, with the total variation distance metric. Therefore, by the contraction mapping theorem on metric spaces, for any initial distribution μ_0, the evolution of the mean field measures, $\mu_0, \mu_1, \mu_2, \ldots$ converges to the unique fixed point of the map $\mathcal{T}_\sigma$. □

In Section 8, we show via numerical experiments that there is reason for optimism: at least in the presence of negative externalities, it appears our bound is likely quite loose. In particular, in these regimes the dynamics continue to converge despite failure of the contraction condition. Establishing a stronger sufficient condition for convergence remains an open question.

7. From the finite agent system to the mean field model In defining the mean field steady state concept in Section 5, an implicit assumption is that the total number of agents is infinite and drawn from a continuum. The return for making this approximation is significant analytical simplification, allowing us to obtain the existence, uniqueness, and convergence results in the preceding sections. In this section we look at the roots of the mean field model, and ask whether it faithfully models the limiting dynamics of finite agent systems as the system size increases. We provide formal guarantees that this approximation property is valid. In particular, we show that under the same contraction condition introduced in the preceding section, the sequence of stochastic systems with finitely many users converges *uniformly* over time to the dynamical system of the mean field model.

Our analysis requires notation pertaining two different systems. The first is a sequence of systems with finitely many agents, indexed by the number of agents m; all quantities relative to the finite agent system will be labeled with a superscript m. In particular, μ_t^m denotes the (random) occupation measure at time t in the finite agent system with m agents. The second is the mean field dynamical system, as introduced in Section 3 and studied in the preceding sections. In particular, μ_t denotes the (deterministic) occupation measure at time t in this dynamical system.

Our main result is an approximation theorem that demonstrates that the dynamics of the finite system converge uniformly over time to the trajectory of the mean field system. To state the result, we require the following definitions. Let $\mathcal{B}_n \subset [0,1]^n$ be the set of all *boxes* in n dimensions, i.e.:

$$\mathcal{B}_n = \left\{ \prod_{i=1}^{n} [a_i, b_i] : 0 \leq a_i \leq b_i \leq 1 \right\}.$$

Given two probability measures μ and ν on $\mathbb{Z}_+^{2n} \times [0,1]^n$, define a distance measure ρ between them as follows:

$$\rho(\mu, \nu) = \sup_{\boldsymbol{z} \in \mathbb{Z}_+^{2n}, B \in \mathcal{B}_n} |\mu(\boldsymbol{z}, B) - \nu(\boldsymbol{z}, B)|. \tag{12}$$

This defines a metric on probability measures; it is straightforward to check that convergence in this metric is weaker than convergence in total variation, but stronger than weak convergence. Our theorem is stated in terms of this distance measure. The class of sets $\mathcal{B}_n$, appearing in the defintion of ρ, has finite Vapnik-Chervonenkis (VC) dimension, and thus admits a uniform law of large numbers; we crucially employ this property in our approximation theorem.

THEOREM 2. *When $\beta(1+L) < 1$, μ_t^m converges to μ_t uniformly over t in the metric ρ:*

$$\lim_{m \to \infty} \sup_{t \geq 0} \mathbb{E}[\rho(\mu_t^m, \mu_t)] = 0. \tag{13}$$

(The expectation in the theorem appears because the empirical measure μ_t^m is random.)

To prove the theorem, we introduce an additional stochastic system that we employ to aid in relating the finite agent systems to the mean field model; we refer to this third system as the *auxiliary* system. In the auxiliary system, agents transition according to the mean field population profile, rather than their empirical population profile. This fact helps in relating it to the mean field system. At the same time, these agents are coupled (in a sense made precise in the proofs below) with the agents in the finite systems. We then show that the auxiliary system converges to the mean field system, and the finite system converges to the auxiliary system, in a sense made precise below.

The auxiliary system consists of a countably infinite collection of fictitious agents, indexed by $k = 1, 2, \ldots$. The type and state of the k'th agent at time t are denoted by $\boldsymbol{\phi}_t(k)$ and $\boldsymbol{v}_t(k)$, respectively. We define ν_t^m as the empirical occupation measure of the *first* m agents in the auxiliary system. Formally, for each state $\boldsymbol{z}$ and $\Theta \subseteq [0,1]^n$:

$$\nu_t^m(\boldsymbol{z}, \Theta) \triangleq \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\boldsymbol{v}_t(k) = \boldsymbol{z}, \boldsymbol{\phi}_t(k) \in \Theta\}.$$

The key shift in the auxiliary system is in the transition dynamics of each agent: agent transitions are now assumed to depend on the mean field population profile. Formally, we assume an agent in the auxiliary system at state $\boldsymbol{z}$ and type $\boldsymbol{\theta}$ transitions independently according to the kernel $\mathbb{K}(\cdot \mid \boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{f}_t, \sigma)$, where we recall that $\mathbb{K}$ was defined in Equation 1. We emphasize that $\boldsymbol{f}_t$ denotes the *deterministic* mean field population profile, rather than the (random) empirical population profile in the m'th system. The initial state is $\boldsymbol{v}_0(k) = 0$ for all k, and the initial type of each agent is sampled independently from W.

Proof of Theorem 2. The proof proceeds by a sequence of lemmas. Lemma 1 first relates the random auxiliary system to the mean field dynamical system defined; its proof is provided in the appendix.

LEMMA 1. *There holds:*

$$\lim_{m \to \infty} \sup_{t \geq 0} \mathbb{E}[\rho(\nu_t^m, \mu_t)] = 0.$$

The next lemma relates the empirical and mean field population profiles to the expected population profile corresponding to the occupation measures μ_t^m and ν_t^m; the proof is also in the appendix.

Lemma 2.

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\sup_{1\leq i\leq n} |f_t^m(i) - p(\sigma, \mu_t^m)(i)|] = 0; \tag{14}$$

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\sup_{1\leq i\leq n} |f_t(i) - p(\sigma, \nu_t^m)(i)|] = 0. \tag{15}$$

The remainder of the proof exploits a specific stochastic coupling between the finite agent and auxiliary systems as follows. At time 0, the initial state of agents in both systems have the same distribution: each agent k starts with $\boldsymbol{z}_0^m(k) = \mathbf{0}$ and $\boldsymbol{v}_0^m(k) = \mathbf{0}$, with $\boldsymbol{\theta}_0^m(k)$ and $\boldsymbol{\phi}_0^m(k)$ sampled i.i.d. from W. At each subsequent time step, we evolve the state of each agent forward using m independent coupled transitions identical to the coupled transitions defined in the proof of Theorem 1, with the exception that the reward probabilities depend on f_t^m (in the finite agent system) and f_t (in the auxiliary system), respectively; we omit the details. This completely defines the joint distribution of the sequences of empirical measures $\{\mu_t^m\}$ (finite agent system) and $\{\nu_t^m\}$ (auxiliary system). Let $\mathcal{F}_t$ denote the σ-algebra generated by all random variables involved up to time t.

Now note that $\rho(\mu_t^m, \nu_t^m) \leq d_{TV}(\mu_t^m, \nu_t^m)$, by characterization D2 of total variation distance. Using this fact and Lemma 1, we obtain

$$\begin{aligned}
\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\rho(\mu_t^m, \mu_t)] &\leq \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\rho(\mu_t^m, \nu_t^m)] + \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\rho(\nu_t^m, \mu_t)] \\
&= \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\rho(\mu_t^m, \nu_t^m)] \qquad \text{by Lemma 1} \\
&\leq \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[d_{TV}(\mu_t^m, \nu_t^m)]. \qquad (16)
\end{aligned}$$

We proceed by showing a stochastic contraction condition between the coupled versions of the auxiliary and finite agent systems. This stochastic contraction condition suffices to allow us to show the finite and auxiliary systems converge to each other uniformly over time.

The stochastic contraction condition we develop requires a definition of a particular distance measure between the (coupled) empirical measures μ_t^m and ν_t^m. Specifically, we define d_t^m as follows:

$$d_t^m \triangleq \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\left\{(\boldsymbol{z}_t^m(k), \boldsymbol{\theta}_t^m(k)) \neq (\boldsymbol{v}_t^m(k), \boldsymbol{\phi}_t^m(k))\right\}.$$

Note that d_t^m is a $\mathcal{F}_t$-measurable random variable. Also, observe that by characterization D3 of the total variation distance, it follows that:

$$d_{TV}(\mu_t^m, \nu_t^m) \leq d_t^m.$$

This can be seen as follows: d_t^m is equivalent to the probability that an agent sampled uniformly at random from $\{1, \ldots, m\}$ has different states in the finite agent system and the auxiliary system. On the other hand, μ_t^m (resp., ν_t^m) is the law of an agent sampled uniformly at random from the finite agent system (resp., the auxiliary system). Thus we can apply characterization D3 to obtain the preceding inequality. Therefore, to show that the right hand side of (16) is zero, it suffices to show that:

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[d_t^m] = 0.$$

We show that d_t^m obeys a contraction condition in expectation. Formally, we show that under the coupling described above, for every $\epsilon > 0$, there exists m_0 such that for all t and any $m \geq m_0$ there holds:

$$\mathbb{E}[d_{t+1}^m] \leq \beta(1 + L)\mathbb{E}[d_t^m] + \epsilon. \tag{17}$$

To see this, note that:

$$\mathbb{E}[d_{t+1}^m \mid \mathcal{F}_t] = \frac{1}{m}\sum_{k=1}^{n} \mathbb{P}\left((\boldsymbol{z}_{t+1}^m(k), \boldsymbol{\theta}_{t+1}^m(k)) \neq (\boldsymbol{v}_{t+1}^m(k), \boldsymbol{\phi}_{t+1}^m(k)) \mid \mathcal{F}_t\right).$$

We can now analyze each of the terms on the right hand side using exactly the same reasoning as the proof of Theorem 1. In particular, given the coupling that we have enforced, let $E_t(k)$ be the event that the state of agent k at time t in the finite agent system is different from the state in the auxiliary system. Note that the k'th term on the right hand side of the preceding expression is $\mathbb{P}(E_{t+1}(k) \mid \mathcal{F}_t)$. Note that $E_{t+1}(k)$ is contained in the following event: there is no regeneration at time t (since the coupling ensures the post-regeneration states are identical); and either $E_t(k)$ holds, or after the transition to time $t+1$, the subsequent states differ. Thus we conclude, using an argument analogous to that leading to (10):

$$\mathbb{P}(E_{t+1}(k) \mid \mathcal{F}_t) \leq \beta\left(\mathbb{1}_{E_t(k)} + \mathbb{1}_{E_t(k)^c} L \sup_i |f_t^m(i) - f_t(i)|\right).$$

Taking expectations, we conclude:

$$\mathbb{E}[d_{t+1}^m] \leq \beta\left(\mathbb{E}[d_t^m] + L\mathbb{E}\left[\sup_i |f_t^m(i) - f_t(i)|\right]\right). \tag{18}$$

We now use Lemma 2 to bound the second term on the right hand side. In particular, observe that (as shown in the proof of Lemma 2), we have:

$$p(\sigma, \mu_t^m) = \frac{1}{m}\sum_{k=1}^{m} \sigma(\boldsymbol{z}_t^m(k), i).$$

Similarly:

$$p(\sigma, \nu_t^m) = \frac{1}{m}\sum_{k=1}^{m} \sigma(\boldsymbol{v}_t^m(k), i).$$

Thus we have:

$$|p(\sigma, \mu_t^m) - p(\sigma, \nu_t^m)| \leq \frac{1}{m}\sum_{k=1}^{m} |\sigma(\boldsymbol{z}_t^m(k), i) - \sigma(\boldsymbol{v}_t^m(k), i)|$$

Now note that under $E_t(k)^c$, the k'th term on the right hand side is zero; and under $E_t(k)$, the k'th term on the right hand side can be at most 1 (since $0 \leq \sigma(\cdot) \leq 1$). Thus in particular the right hand side is less than or equal to d_t^m, i.e.:

$$|p(\sigma, \mu_t^m) - p(\sigma, \nu_t^m)| \leq d_t^m.$$

This bound can be used together with the two bounds in Lemma 2 to conclude that there exists m_0 such that for $m > m_0$ and for all t, we have

$$\mathbb{E}\left[\sup_i |f_t^m(i) - f_t(i)|\right] \leq d_t^m + \frac{\epsilon}{\beta L}.$$

Combining this with (18) gives (17).

To complete the proof of the theorem, we show that if $\beta(1+L) < 1$, then:

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[d_t^m] = 0. \tag{19}$$

This follows from (17). Fix $\epsilon > 0$ and consider $m > m_0$ as given in the lemma. Since $d_0^m = 0$, by induction we have:

$$\mathbb{E}[d_t^m] \leq (1 + \beta(1+L) + \cdots + \beta^t(1+L)^t)\epsilon.$$

It follows that when $\beta(1+L) < 1$:

$$\sup_t \mathbb{E}\left[d_t^m\right] \leq \frac{\epsilon}{1-\beta(1+L)}.$$

Since ϵ was arbitrary, it follows that $\sup_{t\geq 0} \mathbb{E}[d_t^m] \to 0$ as $m \to \infty$, as required. Returning to (16), we conclude that if $\beta(1+L) < 1$, then

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[\rho(\mu_t^m, \mu_t)] \leq \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[d_{TV}(\mu_t^m, \nu_t^m)] \leq \lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[d_t^m] = 0,$$

which completes the proof. □

8. Computational analysis of dynamics In this section we explore the empirical evolution of the population profile for a variety of reward models, where we simulate the actions of a finite number of agents. In the theoretical analysis above, we showed that when the contraction condition $\beta(1+L) < 1$ holds, the bandit game dynamics converge to its corresponding unique MFSS. In this section, we focus on scenarios where this contraction condition is violated, and show the interesting patterns that emerge under different scenarios. We consider reward probability functions Q with both positive and negative externalities. With negative externalities, the system appears to converge to an MFSS, even when $\beta(1+L) \gg 1$. On the other hand, under positive externalities, this is not the case; instead, we visualize the complex behavior that depends on the parameters of the problem. These simulations then complement the theory: we demonstrate the boundaries of the theorems, while also giving insight into the dynamic behavior in scenarios where the theory does not apply.[1]

8.1. Setup of simulations In the examples considered, agents choose between two arms as that is convenient to visualize. We compute and plot the dynamics of the population profile for various examples of reward models.

Each agent uses a simple upper confidence bound (UCB) strategy (cf. [9]) that tries to maximize rewards. Initially, the agent pulls a random arm that has not been pulled before until all arms are pulled at least once. Thereafter, agents choose uniformly among arms i that maximize the following score:

$$\frac{w_i}{w_i + l_i} + \sqrt{\frac{\log(t)}{w_i + l_i}},$$

where t denotes the number of steps after regeneration and w_i and l_i are the number of wins and losses for arm i, respectively. That is, the agent selects the arm uniformly from those actions with the highest upper confidence bound.[2]

For each parameter setting, we replicate the simulation four times, leading to four different empirical evolution paths, each of which we plot with a different color; these paths are representative of patterns seen over many simulation runs. This is helpful in visualizing convergence behavior (or lack thereof). We show both the scatterplots and a smoothed dynamics path computed based on LOWESS. This way, we can clearly visualize the trend and also the variation around that trend. Finally, besides the continuation probability β, we also report the corresponding average lifespan $T = (1-\beta)^{-1}$, as the latter is more interpretable with respect to the learning rate.

[1] The code for these simulations can be found at https://github.com/schmit/mab_games_sims, where we also provide implementations for ϵ-greedy and Thompson sampling strategies.

[2] We draw these actions randomly in case of ties to avoid artifacts due to indexing, e.g., the first arm being pulled more often simply because it has the lowest index.

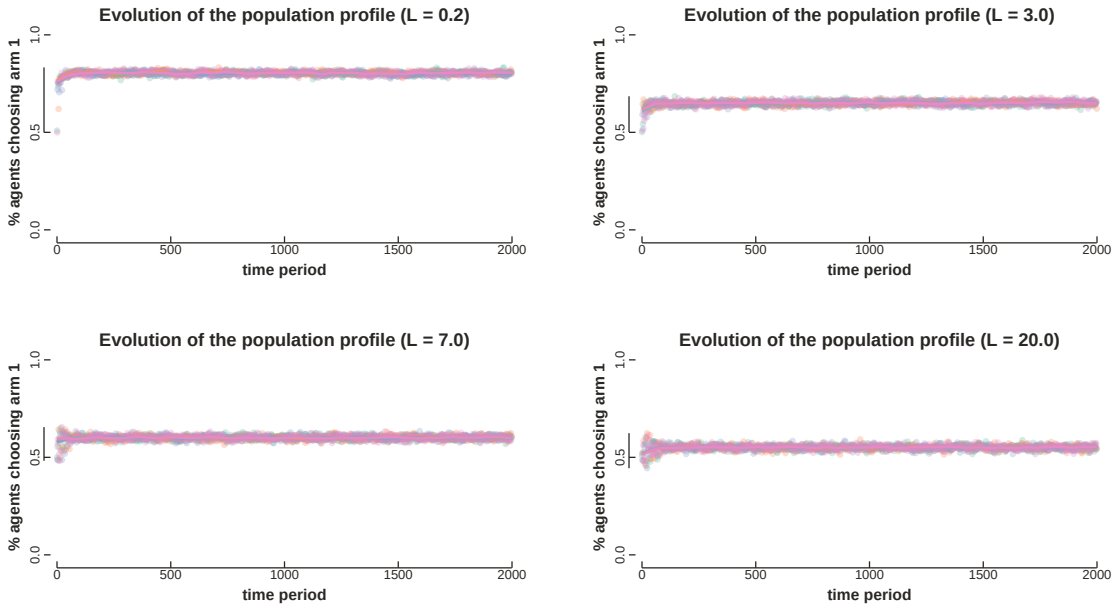

FIGURE 1. Negative externalities: $Q(\theta, f) = \frac{\theta}{1+Lf}$ for agents with average lifespan $T = 50$ ($\beta = 0.98$). Distribution over types is $\theta_1 \sim \text{Beta}(3,1)$ and $\theta_2 \sim \text{Beta}(1,3)$ (independent). Number of agents is 2000.

8.2. Results In this subsection we discuss the results of simulations under different payoff scenarios, and a variety of parameters. We first discuss a payoff function that imposes negative externalities. Thereafter, we consider two different scenarios with positive externalities.

Negative externalities. First, we consider the reward model with $Q(\theta, f) = \theta/(1+Lf)$ while varying L (which corresponds exactly to the Lipschitz constant). Hence, as a larger fraction of agents pulls an arm, the reward probability of that arm decreases. We are interested in the behavior of agents as the we move away from the contraction condition of Theorem 1 that ensures convergence to the MFSS. Unlike in the other experiments, θ drawn uniformly over $[0,1]^2$ is uninteresting in this scenario. We therefore use the following (independent) beta distributions for the types:

$$\theta_1 \sim \text{Beta}(3,1), \quad \theta_2 \sim \text{Beta}(1,3).$$

Hence, a majority of agents would prefer action 1.

To investigate the convergence behavior as we move away from the contraction condition, we initially vary L while keeping the number of agents constant at 2000. In Figure 1 we plot the population dynamics corresponding to different values of L. In these simulations, the average lifespan for agents is $T = 50$, which is equivalent to $\beta = 0.98$. The system appears to converge in every case, regardless of the value of L. As expected, as L increases, the fraction of pulls for action 1 decreases.

To further demonstrate the convergence to the MFSS as the number of agents increases, we plot the agent dynamics of the worst case we simulated in Figure 2. Here, the Lipschitz constant is $L = 20$ and the average lifespan is $T = 200$, or equivalently $\beta = 0.995$. Note that this implies $\beta(1+L) \approx 20 \gg 1$. As the number of agents increases, we indeed observe that the population profile evolution converges to the MFSS. As expected, in cases where the average lifespan is shorter and the Lipschitz constant is smaller, we observe the same outcome; we omit these simulations for brevity.

Positive externalities. Now we change the setting to reward functions that correspond to positive externalities. Contrary to the previous example, the system does not necessarily converge to an

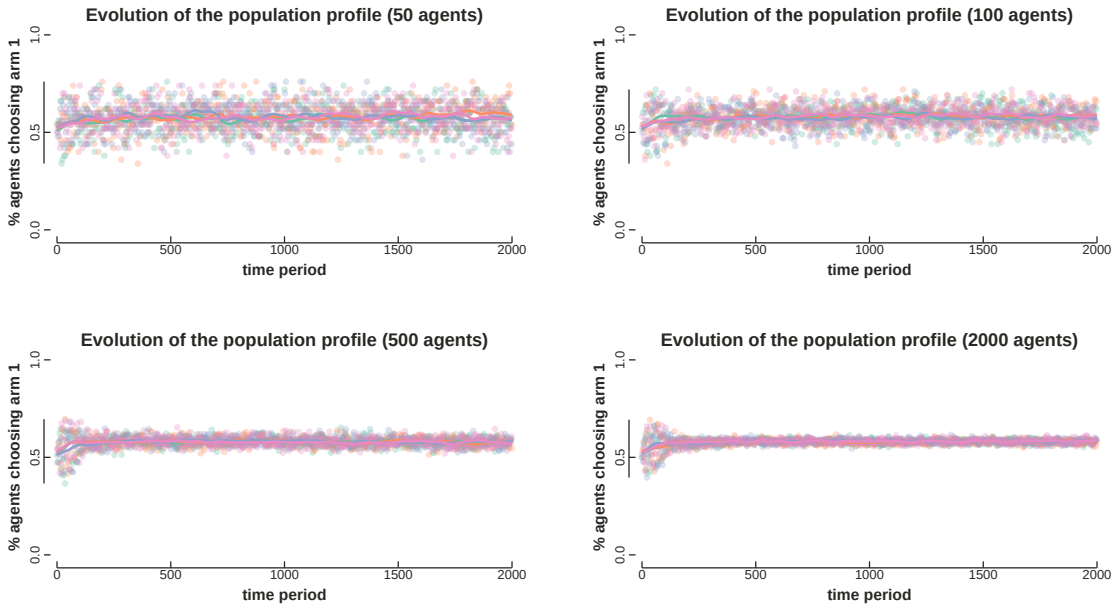

FIGURE 2. Negative externalities: $Q(\theta, f) = \frac{\theta}{1+Lf}$, where $L = 20$, with varying number of agents. We use agents with average lifespan $T = 200$ ($\beta = 0.995$). Distribution over types is $\theta_1 \sim \text{Beta}(3,1)$ and $\theta_2 \sim \text{Beta}(1,3)$ (independent).

MFSS in this setting. We highlight the different dynamics by changing the average lifespan of agents, and by changing the number of agents, both of which affect the empirical dynamics.

The first reward probability function with positive externalities we consider is given by $Q(\theta, f) = \theta f$. We let the distribution of types θ be uniform over $[0,1]^2$. This means that at the population level equal numbers of agents prefer the two arms, though individual agents prefer one over the other.

From theory and simulations we know that the dynamics depend strongly on the lifespan, so first we compare two scenarios with different average lifespans. In the first, the average lifespan of agents is long, while in the second it is short. These scenarios lead to distinctly different dynamics.

Figure 3 shows the dynamics for different numbers of agents when the average lifespan is long: $T = 200$ ($\beta = 0.995$). We observe that the process does not converge to a unique equilibrium; the majority of agents either favors arm 1 or they favor arm 2. Furthermore, when the number of agents increases, switching between these two extremes occurs less frequently. We note that $\beta(1+L) \approx 2 > 1$ in this case, hence this scenario is beyond the scope of the approximation result.

In Figure 4 we show dynamics of the same scenario, except that agents have a much shorter lifespan, with $T = 10$. Here we see that agents do not live long enough to learn to favor one of the actions over the other. As the number of agents increases, the evolution of the profile converges to being stable at 0.5. The condition of the approximation theorem is still violated; $\beta(1+L) = 1.8 > 1$, though not as much as when agents had longer lifetimes of $T = 200$.

We also present another example with positive externalities, namely a type of coordination game with separable rewards. Here, we consider a reward probability function $Q(\theta, f) = \gamma\theta + (1-\gamma)f$. Again, the types are distributed uniformly over $[0,1]^2$. In Figure 5 we plot the evolution of the population profile for $\gamma = 0.5$ while varying the average lifespans of agents. The average lifespans correspond to $\beta = 0.8, 0.95, 0.98$ and 0.995, respectively. We note that if the lifespan is short, the evolution is stationary as agents are unable to learn about the population dynamics. However, for longer average lifespans, we see that the majority of agents prefer one of the arms. This agrees with the previous results.

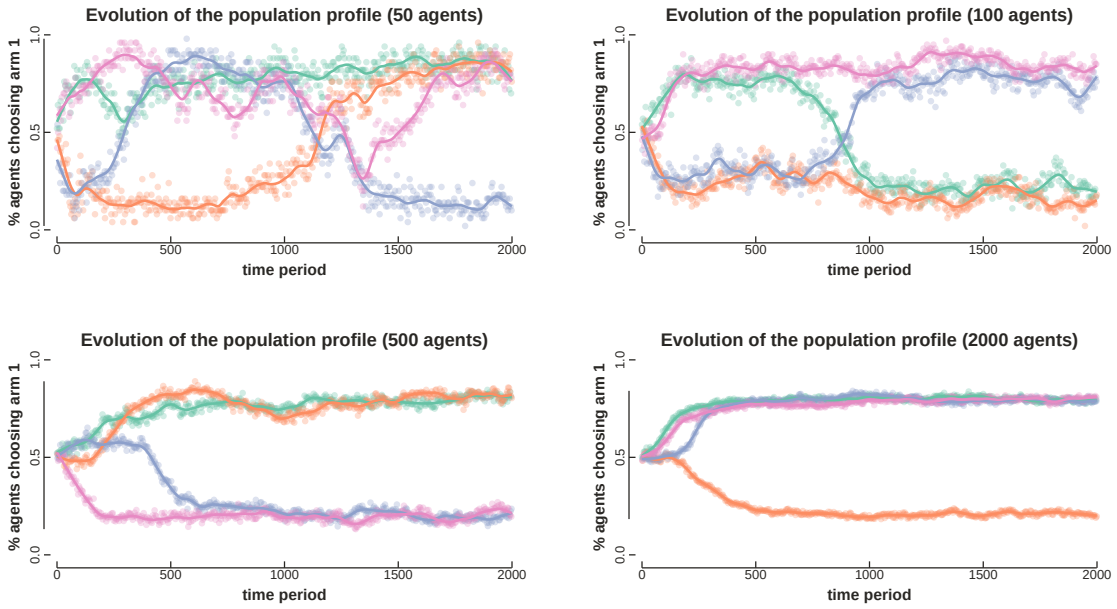

FIGURE 3. Positive externalities: $Q(\theta, f) = \theta f$ with a long average lifespan of $T = 200$ time periods ($\beta = 0.995$). The type distribution is uniform over $[0, 1]^2$.

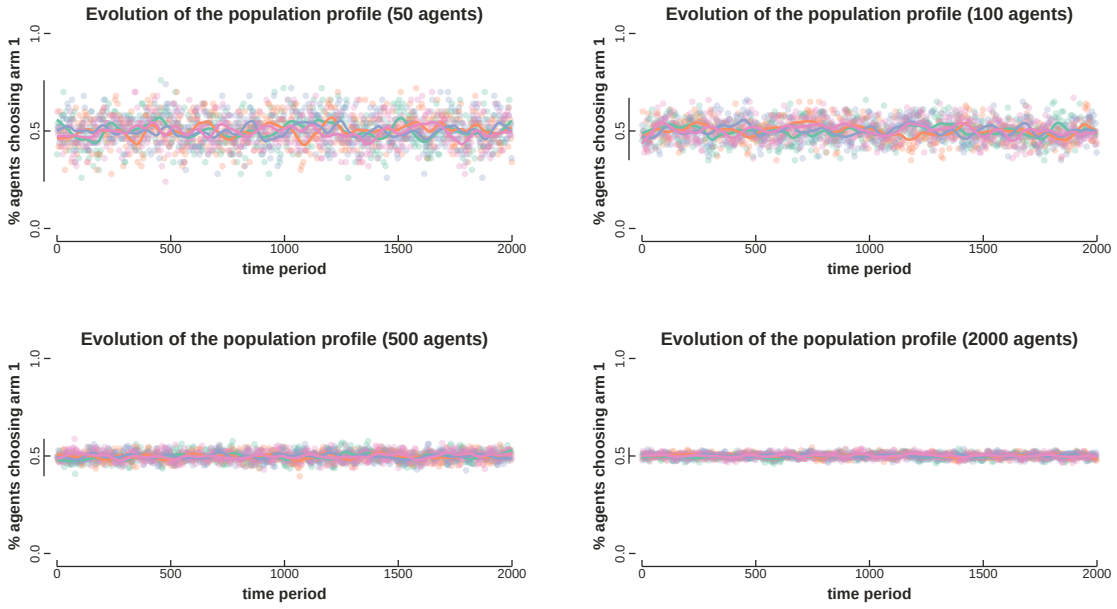

FIGURE 4. Positive externalities: $Q(\theta, f) = \theta f$ with a short average lifespan of $T = 10$ time periods ($\beta = 0.9$). The type distribution is uniform over $[0, 1]^2$.

Taken together, these examples suggest that in cases with positive externalities, our convergence condition is more brittle than in cases with negative externalities. In particular, the effect of long-lived agents in settings with positive externalities is that coordination on specific actions becomes unstable: even if all agents have temporarily converged to playing one action, eventually the pop-

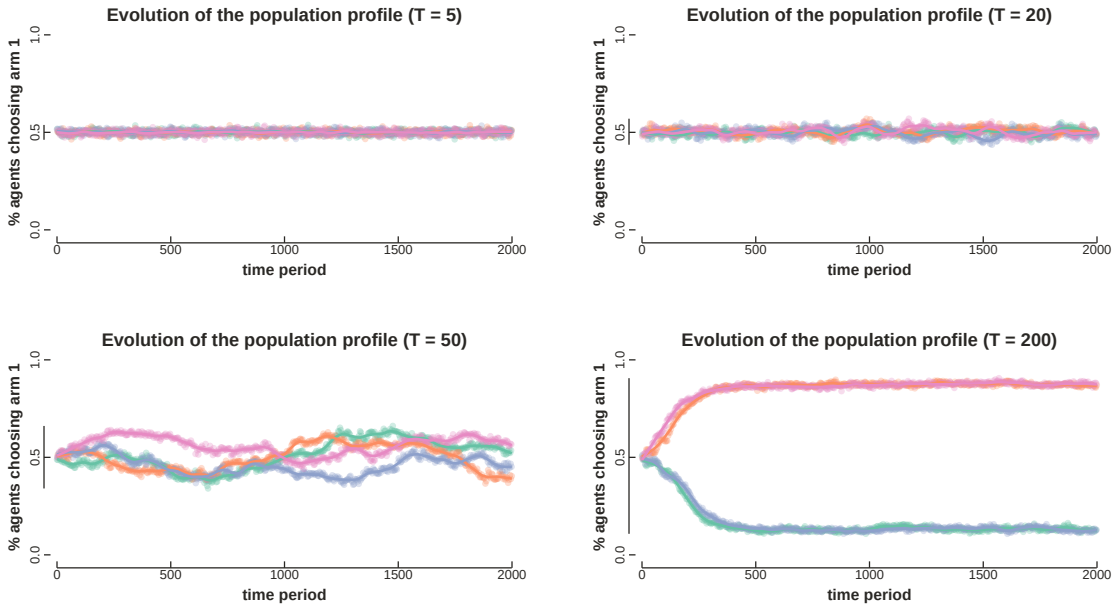

FIGURE 5. Separable rewards: $Q(\theta, f) = \frac{\theta+f}{2}$ for varying average lifespans ($\beta = 0.8, 0.95, 0.98$ and 0.995). The type distribution is uniform over $[0,1]^2$. Number of agents is 2000.

ulation can "drift" to playing another action instead. This is similar in spirit to the instability of Nash equilibria in coordination games (see, e.g., [18]).

9. Conclusion In this paper, we studied the dynamics of multi-armed bandit problems when the agent rewards are dependent on each other through their cumulative 'population profile'. An important insight from our analysis is that, under appropriate assumptions, the system reaches a unique steady state. One interpretation of this result is that it provides justification for the use of MAB policies optimized for stationary environments, even when the environment may be nonstationary due to the actions of others.

At the same time, our work leaves several open questions. Notably, our results do not make any significant assumptions about either the policies used or the nature of the reward function; as a result, the condition (6) obtained for uniqueness and convergence to MFSS is inherently conservative. If one considers restrictions on both the type of reward and the policies used, we might aim to obtain a similar result under weaker assumptions on either the regeneration factor β or the Lipschitz constant L (as suggested by our numerical experiments). One example of a result along these lines is presented in Appendix C, where we show uniqueness of MFSS for any decreasing reward function, under an appropriate assumption on the policies used. Proving convergence to MFSS in such a setting remains an open direction.

Appendix A: Proofs

A.1. Proposition 1 We establish the result through a sequence of lemmas. Our approach takes three steps. First, we show that for a given policy and fixed population profile, the resulting steady state distribution of the agent dynamics is continuous in the population profile. Next, we use this fact to show that if we fix the policy and a population profile, and compute the new population profile induced by the resulting steady state distribution, that map is continuous in the initial population profile. The proof is completed by appealing to Brouwer's fixed point theorem

(e.g. [27]). Throughout the proof we endow the space of measures $\mathcal{M}$ with the topology of weak convergence.

In the following lemma, we let $P_s(\boldsymbol{z}'|\boldsymbol{\theta},\boldsymbol{f},\sigma)$ denote the conditional distribution of the state of the agent given that the last regeneration occurred s time steps in the past, for a fixed $\boldsymbol{\theta},\boldsymbol{f}$ (over all time), and σ; formally, we have inductively:

$$P_0(\boldsymbol{z}'|\boldsymbol{\theta},\boldsymbol{f},\sigma) = \mathbb{1}_{\boldsymbol{z}=\boldsymbol{0}}; \tag{20}$$

$$P_s(\boldsymbol{z}'|\boldsymbol{\theta},\boldsymbol{f},\sigma) = \sum_{\boldsymbol{z}} P_{s-1}(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f},\sigma)\sum_{i=1}^{n}\sigma(\boldsymbol{z},i)\Big(Q(\theta(i),f(i))\mathbb{1}_{\boldsymbol{z}'=\boldsymbol{z}+\boldsymbol{w}_i} \tag{21}$$

$$+(1-Q(\theta(i),f(i)))\mathbb{1}_{\boldsymbol{z}'=\boldsymbol{z}+\boldsymbol{l}_i}\Big), \quad s>0. \tag{22}$$

We have the following lemma.

LEMMA 3. *For fixed $\boldsymbol{z},\boldsymbol{\theta}$, and σ, $P_s(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f},\sigma)$ is continuous in $\boldsymbol{f}$.*

Proof. The proof follows by induction. P_0 is trivially continuous in $\boldsymbol{f}$. Each P_s is then continuous, because the sum on the right hand side of (21) is finite (at most finitely many states can be reached in s time steps, starting from $\boldsymbol{z}=\boldsymbol{0}$); and $Q(\theta,f)$ is continuous in f for every θ. □

Next, we show that there is a unique occupation measure that satisfies C1, given a policy and a population profile.

LEMMA 4. *For any pair of a policy σ and population profile $\boldsymbol{f}$, there exists a unique distribution $\Phi(\sigma,\boldsymbol{f})$ that satisfies C1. Further, for fixed σ, $\Phi(\sigma,\boldsymbol{f})$ is continuous in $\boldsymbol{f}$.*

Proof. With $(\sigma,\boldsymbol{f})$ fixed, the kernel $\mathbb{K}\left(\cdot\,\middle|\,\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{f},\sigma\right)$ represents a Markov chain with state space $\mathbb{Z}_+^{2n}\times[0,1]^n$. Due to the independent regeneration with geometric intervals, the expression for its invariant distribution, $\Phi(\sigma,\boldsymbol{f})$ can be explicitly written by conditioning on the regeneration interval as follows:

$$\begin{aligned}\Phi(\sigma,\boldsymbol{f})(\boldsymbol{z},\Theta) &= \int_{\boldsymbol{\theta}\in\Theta}\sum_{s\geq 0}(1-\beta)\beta^s P_s(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f},\sigma)\,dW\\ &= \sum_{s\geq 0}(1-\beta)\beta^s\int_{\boldsymbol{\theta}\in\Theta}P_s(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f},\boldsymbol{\sigma})\,dW.\end{aligned}$$

(The second equality follows by the monotone convergence theorem.) Let $\boldsymbol{f}_\ell\to\boldsymbol{f}$ under the Euclidean norm. From Lemma 3, $P_s(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f}_\ell,\sigma)\to P_s(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{f}_\ell,\sigma)$. Using bounded convergence, this implies that $\Phi(\sigma,\boldsymbol{f}_\ell)(\boldsymbol{z},\Theta)\to\Phi(\sigma,\boldsymbol{f})(\boldsymbol{z},\Theta)$ for all states $\boldsymbol{z}$ and Borel sets Θ. Therefore, $\Phi(\sigma,\boldsymbol{f}_\ell)\to\Phi(\sigma,\boldsymbol{f})$ weakly. □

Now recalling the definition of p in (3), define $F_\sigma: S\mapsto S$ by:

$$F_\sigma(\boldsymbol{f}) \triangleq p(\sigma,\Phi(\sigma,\boldsymbol{f})).$$

Hence F_σ maps population profile $\boldsymbol{f}$ to the population profile induced by $\boldsymbol{f}$ and σ.

LEMMA 5. *F_σ is continuous (under the Euclidean norm).*

Proof. Consider a sequence $\boldsymbol{f}_\ell\to\boldsymbol{f}$ under the Euclidean norm on S. From Lemma 4, we have $\Phi(\sigma,\boldsymbol{f}_\ell)\to\Phi(\sigma,\boldsymbol{f})$ weakly. Since $\mathbb{Z}_+^{2n}\times[0,1]^n$ is separable, the Skorokhod representation theorem implies that we have random variables $(\boldsymbol{z}_\ell,\boldsymbol{\theta}_\ell)$ with corresponding laws $\Phi(\sigma,\boldsymbol{f}_\ell)$, as well as a random variable $(\boldsymbol{z},\boldsymbol{\theta})$ with corresponding law $\Phi(\sigma,\boldsymbol{f})$ such that $(\boldsymbol{z}_\ell,\boldsymbol{\theta}_\ell)\to(\boldsymbol{z},\boldsymbol{\theta})$ almost surely. Since $\boldsymbol{z}_\ell$ lies in a discrete space, this implies that there exists ℓ' such that for all $\ell>\ell'$, $\boldsymbol{z}_\ell=\boldsymbol{z}$. Therefore

for any i, $\sigma(\boldsymbol{z}_\ell, i) \to \sigma(\boldsymbol{z}, i)$ almost surely. In particular, by the bounded convergence theorem we have:

$$\begin{aligned} F_\sigma(\boldsymbol{f}_\ell)(i) &= p(\sigma, \Phi(\sigma, \boldsymbol{f}_\ell))(i) \\ &= \mathbb{E}[\sigma(\boldsymbol{z}_\ell, i)] \to \mathbb{E}[\sigma(\boldsymbol{z}, i)] \\ &= p(\sigma, \Phi(\sigma, \boldsymbol{f}))(i) = F_\sigma(\boldsymbol{f})(i), \end{aligned}$$

as required. □

We can now prove that there exists an MFSS.

Proof of Proposition 1. Since F_σ is continuous and maps a compact set S to itself, using Brouwer's theorem, there exists a fixed point, $\boldsymbol{f}^*$ satisfying $F_\sigma(\boldsymbol{f}^*) = \boldsymbol{f}^*$. Then it can be verified that the pair $(\Phi(\sigma, \boldsymbol{f}^*), \boldsymbol{f}^*)$ satisfies both C1 and C2 simultaneously. □

A.2. Lemmas 1 and 2

Proof of Lemma 1. Consider any agent k, from the auxiliary system. For this agent $(\boldsymbol{v}_0(k), \boldsymbol{\phi}_0(k))$ is sampled from μ_0. Suppose $(\boldsymbol{v}_t(k), \boldsymbol{\phi}_t(k))$ have the same distribution as μ_t. Since $\boldsymbol{f}_t = p(\sigma, \mu_t)$, the transition kernel, $\mathbb{K}\left((\boldsymbol{z}, \Theta) \middle| (\boldsymbol{z}', \boldsymbol{\theta}); \boldsymbol{f}_t, \sigma\right)$, as we defined for the given agent corresponds exactly to the map T_σ. Hence, the distribution of $(\boldsymbol{v}_{t+1}(k), \boldsymbol{\phi}_{t+1}(k))$ is given by μ_{t+1}. Therefore by induction, for each t, the probability distribution of $(\boldsymbol{v}_t(k), \boldsymbol{\phi}_t(k))$ is given by μ_t. Furthermore, the transition kernels are all independent across the agents, since $\boldsymbol{f}_t$ is deterministic for every t. Therefore, the occupation measure ν_t^m is equal to the occupation measure of a sample of m independently drawn agents from the distribution μ_t for every t. We conclude that for each t, $\{\mathbb{1}\{\boldsymbol{v}_t(k) = \boldsymbol{z}, \boldsymbol{\phi}_t(k) \in \Theta\} : 1 \le k \le m\}$ is a collection of m independent Bernoulli random variables with non-zero probability of success $\mu_t(\boldsymbol{z}, \Theta)$.

To complete the result, we now employ standard results on uniform laws of large numbers [14]. The following concentration bound—the *Vapnik-Chervonenkis (VC) inequality*—holds for all $\boldsymbol{z}$, m, and $\epsilon > 0$:

$$\mathbb{P}\left(\sup_{\boldsymbol{z}, B \in \mathcal{B}_n} |\nu_t^m(\boldsymbol{z}, B) - \mu_t(\boldsymbol{z}, B)| > \epsilon\right) \le h(m) \triangleq 8\mathcal{S}(\mathcal{C}, m) e^{-m\epsilon^2/32},$$

where $\mathcal{S}(\mathcal{C}, m)$ is the *shattering coefficient* of the following class:

$$\mathcal{C} = \{\{\boldsymbol{z}\} \times B : \boldsymbol{z} \in \mathbb{Z}_+^{2n}, B \in \mathcal{B}_n\}.$$

When the Vapnik-Chervonenkis (VC) dimension of a class is finite, the shattering coefficient grows at most polynomially in m, and thus $h(m) \to 0$ as $m \to \infty$. Thus it suffices to show that $\mathcal{C}$ has finite VC dimension.

The class $\mathcal{C}$ can be viewed as the product of two classes of sets: $\mathcal{A} = \{\{\boldsymbol{z}\} : \boldsymbol{z} \in \mathbb{Z}_+^{2n}\}$, and $\mathcal{B}_n$. The Vapnik-Chervonenkis dimension (VC) of the class $\mathcal{A}$ is finite (in fact equal to one, since it is a collection of singletons). The VC dimension of the class $\mathcal{B}_n$ is finite (in fact equal to $2n$). The product of two classes with finite VC dimension is finite (cf. Theorem 9.2.6 of [15]), and thus $\mathcal{C}$ has finite VC dimension.

To complete the proof, since $|\nu_t^m(\boldsymbol{z}, B) - \mu_t(\boldsymbol{z}, B)| \le 1$, it follows that:

$$\mathbb{E}[\rho(\nu_t^m, \mu_t)] \le (1 - h(m))\epsilon + \epsilon.$$

The right hand side does not depend on t and approaches 2ϵ as $m \to \infty$. Therefore:

$$\lim_{m \to \infty} \sup_{t \ge 0} \mathbb{E}[\rho(\nu_t^m, \mu_t)] \le 2\epsilon.$$

Since ϵ was arbitrary, the result follows. □

Proof of Lemma 2. To prove equation (14), we first write:

$$f_t^m(i) = \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\{\sigma_t(k, \boldsymbol{z}_t^m(k)) = i\}.$$

Recall that:

$$\begin{aligned} p(\sigma, \mu_t^m)(i) &= \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(\boldsymbol{z}, i)\mu_t^m(\boldsymbol{z}, [0,1]^n) \\ &= \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(\boldsymbol{z}, i)\frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\{\boldsymbol{z}_t^m(k) = \boldsymbol{z}\} \\ &= \frac{1}{m}\sum_{k=1}^{m} \sigma(\boldsymbol{z}_t^m(k), i). \end{aligned}$$

Therefore:

$$f_t^m(i) - p(\sigma, \mu_t^m)(i) = \frac{1}{m}\sum_{k=1}^{m} \big(\mathbb{1}\{\sigma_t(k, \boldsymbol{z}_t^m(k)) = i\} - \sigma(\boldsymbol{z}_t^m(k), i)\big).$$

Note that the k'th indicator term inside the summation is a Bernoulli random variable, with mean $\sigma(\boldsymbol{z}_t^m(k), i)$; further, these terms are independent across k. Since Bernoulli random variables have variance no larger than $1/4$, it follows that:

$$\mathrm{Var}(f_t^m(i) - p(\sigma, \mu_t^m)(i)) \leq \frac{1}{4m}.$$

It follows by Chebyshev's inequality that:

$$\mathbb{P}(|f_t^m(i) - p(\sigma, \mu_t^m)(i)| > \epsilon) \leq \frac{1}{4m\epsilon^2}.$$

The right hand side is independent of t, and since $|f_t^m(i) - p(\sigma, \mu_t^m)(i)|$ cannot be larger than one, we obtain:

$$\sup_{t\geq 0} \mathbb{E}[|f_t^m(i) - p(\sigma, \mu_t^m)(i)|] \leq \epsilon + \frac{1}{4m\epsilon^2},$$

Taking $m \to \infty$, we find that:

$$\lim_{m\to\infty} \sup_{t\geq 0} \mathbb{E}[|f_t^m(i) - p(\sigma, \mu_t^m)(i)| \leq \epsilon.$$

Since ϵ is arbitrary, and there are only finitely many actions i, (14) follows.

We now show equation (15) making use of Lemma 1. We have:

$$\begin{aligned} \mathbb{E}[|f_t(i) - p(\sigma, \nu_t^m)(i)|] &= \mathbb{E}\left[\left|\sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \big(\mu_t(\boldsymbol{z}, [0,1]^n) - \nu_t^m(\boldsymbol{z}, [0,1]^n)\big)\sigma(\boldsymbol{z}, i)\right|\right] \\ &\leq \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(\boldsymbol{z}, i)\mathbb{E}[|\mu_t(\boldsymbol{z}, [0,1]^n) - \nu_t^m(\boldsymbol{z}, [0,1]^n|)]. \end{aligned} \tag{23}$$

Now note that since each agent regenerates at each period with parameter $1-\beta < 1$, we can choose $K_\epsilon > 0$ such that the probability an agent survives for more than K_ϵ periods without regenerating is at most ϵ; specifically we can choose $K_\epsilon > \log\epsilon/\log\beta$. Define the (finite) set Z_ϵ by:

$$Z_\epsilon = \{\boldsymbol{z} \in \mathbb{Z}_+^{2n} : \|\boldsymbol{z}\|_1 \leq K_\epsilon\}.$$

It follows that $\mu_t(Z_\epsilon^c, [0,1]^n) \leq \epsilon$ for all t; in other words, the measures $\{\mu_t\}_{t\geq 0}$ are tight. Recall that $\mathbb{E}[\nu_t^m(\boldsymbol{z}, B)] = \mu_t(\boldsymbol{z}, B)$ (cf. the proof of Lemma 1), so we have $\mathbb{E}[\nu_t(Z_\epsilon^c, [0,1]^n)] \leq \epsilon$ for all t as well.

Combining this with (23), we obtain:

$$\sup_{t\geq 0} \mathbb{E}[|f_t(i) - p(\sigma, \nu_t^m)(i)|] \leq \sum_{\boldsymbol{z}\in Z_\epsilon} \sigma(\boldsymbol{z}, i) \sup_{t\geq 0} \mathbb{E}[|\mu_t(\boldsymbol{z}, [0,1]^n) - \nu_t^m(\boldsymbol{z}, [0,1]^n)|] + 2\epsilon.$$

Since Z_ϵ is a finite set, the right hand side of the preceding expression approaches 2ϵ as $m \to \infty$, by Lemma 1. Since ϵ is arbitrary, and there are only finitely many actions i, (15) follows. □

Appendix B: Extension to heterogeneous policy spaces In this section, we derive a generalization of the original model that relaxes the assumption that all agents adopt an identical policy σ. Instead, we assume that an agent chooses its policy from a class Σ, independently of all other agents. This is modeled by a probability distribution D on Σ, where $D(\sigma)$ represents the probability of an agent to adopt policy $\sigma \in \Sigma$. The following results describe the formal generalization to analogous results even for heterogeneous policies across the agent population. In order to extend the notion of mean field equilibrium to this setting, we now need to consider a more general measure, μ, that is jointly defined over the agent types, states as well as policies. This requires a modification to the interpretation of the the transition kernel $\mathbb{K}$, which is now relaxed to allow for considering the policy as an additional parameter in Equation (1). In place of Definition 1 from Section 3, an MFSS is now defined as follows.

Definition 3. Given a type distribution W, a reward function Q, and a distribution D over a finite[3] policy space Σ, a *mean field steady state* (MFSS) is defined as a pair $(\mu, \boldsymbol{f})$ where $\mu \in \mathcal{M}$ is a joint distribution on state, type and policy space, $\mathbb{Z}_+^{2n} \times [0,1]^n \times \Sigma$, and $\boldsymbol{f}$ is a population profile, such that the following three conditions hold:

C1 *Stationarity*: For any state $\boldsymbol{z}$, type set Θ and policy $\sigma \in \Sigma$:

$$\mu(\boldsymbol{z}, \Theta, \sigma) = \int_{\boldsymbol{z}', \boldsymbol{\theta}} \mathbb{K}_{\sigma'}\left(\boldsymbol{z}, \Theta \middle| \boldsymbol{z}', \theta; \boldsymbol{f}, \sigma'\right) \mathbb{1}_{\sigma=\sigma'} d\mu_h;$$

C2 *Consistency*: For all i, $1 \leq i \leq n$,

$$f(i) = \sum_{\boldsymbol{z}, \sigma} \sigma(\boldsymbol{z}, i) \mu(\boldsymbol{z}, [0,1]^n, \sigma).$$

C3 μ_h has a marginal on Σ that matches the given distribution D over the policy space Σ.

The mean field dynamics are correspondingly defined through the modified operator $\mathcal{T}_\Sigma$:

$$\mathcal{T}_\Sigma(\mu_h)(\boldsymbol{z}, \Theta, \sigma) \triangleq \int_{(\boldsymbol{z}', \boldsymbol{\theta}, \sigma')} \mathbb{K}_{\sigma'}\left(\boldsymbol{z}, \Theta \middle| \boldsymbol{z}', \boldsymbol{\theta}; \boldsymbol{f}, \sigma'\right) \mathbb{1}_{\sigma=\sigma'} d\mu_h. \tag{24}$$

With these definitions in place, the existence, uniqueness, and approximation results all generalize using a proof approach similar to the homogeneous case. The proof argument effectively involves generalization of the arguments to also "average" over the randomness in the policy. In particular, finiteness of the policy space ensures that the concentration arguments we employ for the approximation theorem Theorem 2 remain essentially unchanged, using a simple union bound. We omit the details.

[3] It is possible to generalize this to a compact set of policy spaces, but we describe the extension assuming a finite set. This keeps notation simple by avoiding the necessity of having to refer to measures on subsets of the policy space.

Appendix C: Negative externalities and uniqueness of MFSS Our previous results apply to a very broad class of policies that only depend on the state vector, and derive conditions on the model that can assure us of uniqueness of equilibria. However, as we found in Section 8, the condition of Theorem 1 is particularly loose in settings with negative externalities. This motivates us to study whether, by making additional assumptions on the policy and reward structure, we can obtain similarly sharp characterizations of MFSS.

In this section, we make progress by restricting our focus to a subclass of policies with *positive sensitivity to arm rewards*, in a sense to be made precise below. We show that MFSS are unique as long as arm rewards are strictly decreasing for every agent type, i.e. $Q(\theta, f)$ is decreasing in f for every θ – a particular form of negative externality. While this is not a strict superset of the conditions under which we derive equilibria in the Section 6, we note that there is no longer a restriction on the sensitivity of the reward function beyond mere continuity, which is needed for existence.

First, we formally define a *bandit process*; this definition is needed to formalize the policy class being considered.

Definition 4 (The bandit process $\boldsymbol{z}_t^{\boldsymbol{R}}$ for a given reward vector $\boldsymbol{R}$ and policy σ). Given a policy σ, let $\boldsymbol{R} \in [0,1]^n$ specify a vector of means corresponding to the Bernoulli random rewards on n arms. Then $\boldsymbol{z}_t^{\boldsymbol{R}}$ denotes the stochastic bandit process on state space $\mathbb{Z}_+^{2n}$, whose evolution is determined by applying the policy σ while facing Bernoulli arm rewards specified by the mean vector $\boldsymbol{R}$, and with no regeneration.

Formally, $\boldsymbol{z}_t^{\boldsymbol{R}}$ is a $2n$-dimensional discrete-time Markov process, with $\boldsymbol{z}_0^{\boldsymbol{R}} = \boldsymbol{0}$, and with $\boldsymbol{z}_{t+1}^{\boldsymbol{R}}$ determined as follows:

1. Let $\hat{\sigma}_t$ be a random variable denoting the arm pulled at time t, with distribution $\sigma(\boldsymbol{z}_t^{\boldsymbol{R}})$.
2. Let Z_t be a Bernoulli random variable with mean $R_{\sigma(\boldsymbol{z}_t^{\boldsymbol{R}})}$.
3. Increment $\boldsymbol{z}_t^{\boldsymbol{R}}$ by one in the coordinate corresponding to a win (if $Z_t = 1$) or loss (if $Z_t = 0$) on arm $\hat{\sigma}_t$.

Positive sensitivity to arm rewards is defined as follows.

Definition 5 (Positive sensitivity to arm rewards (PSAR)). Let $\boldsymbol{a}, \boldsymbol{b} \in (0,1)^n$ be any two *distinct* reward vectors. Denote $A^* \triangleq \{i : a_i \geq b_i\}$, with at least one strict inequality without loss of generality (swapping the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, if necessary). σ satisfies PSAR if the following condition on the corresponding bandit processes holds for all $t \geq 1$:

$$P(\hat{\sigma}(\boldsymbol{z}_t^{\boldsymbol{a}}) \in A^*) > P(\hat{\sigma}(\boldsymbol{z}_t^{\boldsymbol{b}}) \in A^*)$$

The main result of this section shows that there is a unique MFSS when the policy satisfies the condition of Definition 5 and Q is strictly decreasing in f.

Theorem 3. *Suppose σ satisfies PSAR (cf. Definition 5). Further, suppose that the reward function, $Q(\theta, f)$ is strictly decreasing in f for any $\theta \in [0,1]$. Then, there exists a unique MFSS (μ^*, f^*).*

Proof. Since there exists at least one MFSS according to Theorem 1, it is sufficient to show that the existence of two distinct $(\mu^a, \boldsymbol{f}^a)$ and $(\mu^b, \boldsymbol{f}^b)$, both satisfying C1 and C2 leads to a contradiction. To do so, assume $(\mu^a, \boldsymbol{f}^a)$ and $(\mu^b, \boldsymbol{f}^b)$ are a distinct pair of MFSS. Consider the following vector notation for any $\boldsymbol{\theta}, \boldsymbol{f} \in [0,1]^n$:

$$\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{f}) \triangleq (Q(\theta_1, f_1), \ldots, Q(\theta_n, f_n))$$

Arguing similar to the proof of Lemma 4, we have the following relation for any pair $(\mu, \boldsymbol{f})$ that satisfies C1:

$$\mu(\boldsymbol{z}, [0,1]^n) = \int_{\boldsymbol{\theta}} \sum_{t \geq 0} (1-\beta)\beta^t P\left(\boldsymbol{z}_t^{\boldsymbol{Q}(\theta, \boldsymbol{f})} = \boldsymbol{z}\right) \, dW$$

Let $A^* \triangleq \{i : f_i^a \leq f_i^b\}$. Then since Q is strictly decreasing in f, we find that for every $\theta \in [0,1]$ and any $i \in A^*$, there holds $Q(\theta, f_i^a) \geq Q(\theta, f_i^b)$, with strict inequality on at least on one coordinate without loss of generality. Using C2 in the definition of MFSS we then obtain:

$$\begin{aligned}
\sum_{i\in A^*} f_i^a &= \sum_{i\in A^*} \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(\boldsymbol{z}, i)\mu^a(\boldsymbol{z}, [0,1]^n) \\
&= \sum_{i\in A^*} \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(\boldsymbol{z}, i) \int_{\boldsymbol{\theta}} \sum_{t\geq 0} (1-\beta)\beta^t P\left(\boldsymbol{z} = \boldsymbol{z}_t^{\boldsymbol{Q}(\theta, \boldsymbol{f}^a)}\right) dW \\
&= \sum_{t\geq 0} (1-\beta)\beta^t \int_{\boldsymbol{\theta}} \sum_{i\in A^*} \sum_{\boldsymbol{z}\in\mathbb{Z}_+^{2n}} \sigma(z, i) P\left(\boldsymbol{z} = \boldsymbol{z}_t^{\boldsymbol{Q}(\theta, \boldsymbol{f}^a)}\right) dW \\
&= \sum_{t\geq 0} (1-\beta)\beta^t \int_{\boldsymbol{\theta}} P\left(\hat{\sigma}(\boldsymbol{z}_t^{\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{f}^a)}) \in A^*\right) dW \\
&> \sum_{t\geq 0} (1-\beta)\beta^t \int_{\boldsymbol{\theta}} P\left(\hat{\sigma}(\boldsymbol{z}_t^{\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{f}^b)}) \in A^*\right) dW \qquad \text{from Assumption 5} \\
&= \sum_{i\in A^*} f_i^b \qquad \text{by symmetry,}
\end{aligned}$$

which is a contradiction to the definition $A^* \triangleq \{i : f_i^a \leq f_i^b\}$. Therefore, for any policy σ that satisfies Definition 5, there exists a unique pair (μ^*, f^*) that simultaneously satisfy the MFSS conditions C1 and C2, as required. □

References

[1] Adlakha, S., R. Johari. 2010. Mean field equilibrium in dynamic games with complementarities. *CoRR* **abs/1011.5677**.

[2] Adlakha, S., R. Johari, G. Y. Weintraub. 2015. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *J. Economic Theory* **156** 269–316.

[3] Agrawal, R. 1995. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability* **27**(4) 1054–1078.

[4] Agrawal, S., N. Goyal. 2012. Analysis of Thompson Sampling for the multi-armed bandit problem. *Proceedings of the 25th Annual Conference on Learning Theory (COLT)* .

[5] Anantharam, V., P. Varaiya, J. Walrand. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays - part I: I.I.D. rewards. *IEEE Transactions on Automatic Control* **32** 968–976.

[6] Anantharam, V., P. Varaiya, J. Walrand. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays - part II: Markovian rewards. *IEEE Transactions on Automatic Control* **32** 977–982.

[7] Arora, S., E. Hazan, S. Kale. 2012. The Multiplicative Weights Update method: a meta-algorithm and applications. *Theory of Computing* **8** 121–164.

[8] Audibert, J.-Y., R. Munos, C. Szepesvari. 2009. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science* **410** 1876–1902.

[9] Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47** 2–3.

[10] Auer, P., N. Cesa-Bianchi, Y. Freund, , R.E. Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing* **32**(1) 48–77.

[11] Bodoh-Creed, A. 2012. Approximation of large dynamic games. Tech. rep., working paper, Cornell University.

[12] Cesa-Bianchi, N., P. Fischer. 1998. Finite-time regret bounds for the multiarmed bandit problem. *Proceedings of the International Conference on Machine Learning (ICML)*. IEEE, 100–108.

[13] Chapelle, O., L. Li. 2012. An empirical evaluation of Thompson Sampling. *Advances in Neural Information Processing Systems (NIPS)* **24**.

[14] Devroye, L., L. Györfi, G. Lugosi. 2013. *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media.

[15] Dudley, R. M. 1984. A course on empirical processes. *Ecole d'été de Probabilités de Saint-Flour XII-1982*. Springer, 1–142.

[16] Duffie, D., S. Malamud, G. Manso. 2009. Information percolation with equilibrium search dynamics. *Econometrica* **77**(5) 1513–1574.

[17] Friesz, T. L., D. Bernstein, T. E. Smith, R. L. Tobin, B.W.Wie. 1993. A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research* **41**(1) 179–191.

[18] Fudenberg, Drew, David K Levine. 1998. *The theory of learning in games*, vol. 2. MIT press.

[19] Gibbs, A. L., F. E. Su. 2002. On choosing and bounding probability metrics. *International Statistical Review* **70**(3) 419–435.

[20] Gittins, J. C. 1989. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, New York, NY.

[21] Gopalan, A., S. Mannor, Y. Mansour. 2014. Thompson sampling for complex online problems. *ICML*, vol. 14. 100–108.

[22] Hopenhayn, H. A. 1992. Entry, exit and firm dynamics in long run equilibrium. *Econometrica* **60**(5) 1127–1150.

[23] Huang, M., P. Caines, R. Malhame. 2007. Large-population cost-coupled lqg problems with nonuniform agents: Individual-mass behavior and decentralized nash equilibria. *IEEE Transactions on Automatic Control* **52**(9) 1560–1571.

[24] Iyer, K., R. Johari, M. Sundararajan. 2011. Mean field equilibria of dynamic auctions with learning. *SIGECOM*.

[25] J. Audibert and R. Munos. 2011. Introduction to bandits: Algorithms and theory. Https://sites.google.com/site/banditstutorial/.

[26] Jovanovic, B., R.W. Rosenthal. 1988. Anonymous sequential games. *Journal of Mathematical Economics* **17** 77–88.

[27] Karamadian, S. 1977. *Fixed points. Algorithms and applications*. Academic Press.

[28] Kelly, F. P. 1981. Multi-armed bandits with discount factor near one: The bernoulli case. *Annals of Statistics* **9** 987–1001.

[29] Lai, T., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22.

[30] Lasry, J. M., P. L. Lions. 2007. Mean field games. *Japanese Journal of Mathematics* **2** 229–260.

[31] Lindvall, T. 1992. *Lectures on the Coupling Mehtod*. John Wiley & Sons, New York.

[32] Littlestone, N., M. Warmuth. 1994. The weighted majority algorithm. *Information and Computation* **108**(2) 212–261.

[33] Mahajan, A., D. Teneketzis. 2007. *Multi Armed Bandit Problems*. Foundations and Applications of Sensor Management, Springer Verlag.

[34] May, B. C., N. Korda, A. Lee, D. S. Leslie. 2012. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* **13** 2069–2106.

[35] Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58**(5) 527–535.

[36] Russo, D., B. Van Roy. 2014. Learning to optimize via posterior sampling. *CoRR* **abs/1301.2609**.

[37] Scott, S. 2010. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* **26** 639–658.

[38] Weber, R. 1992. On Gittins index for multiarmed bandits. *Annals of Probability* **2** 1024–1033.

[39] Weintraub, G. Y., C. L. Benkard, B. Van Roy. 2008. Markov perfect industry dynamics with many firms. *Econometrica* **76**(6) 1375–1411.

[40] Weintraub, G. Y., C. L. Benkard, B. Van Roy. 2011. Industry dynamics: Foundations for models with an infinite number of firms. *J. Economic Theory* **146** 1965–1994.

[41] Whittle, P. 1980. Multi-armed bandits and Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)* **42** 143–149.

[42] Yang, T., P. Mehta, S. Meyn. 2013. Feedback particle filter with mean-field coupling. *In Proc. of 50th IEEE Conference on Decision and Control* .

[43] Yin, H., P. G. Mehta, S. P. Meyn, U. V. Shanbhag. 2010. Synchronization of coupled oscillators is a game. *Proceedings of the IEEE International Conference on Decision and Control (CDC '10)*. IEEE.