Mean field equilibria of multiarmed bandit games

Ramesh Johari

Stanford University

Joint work with Ramki Gummadi (Stanford University) and Jia Yuan Yu (IBM Research)

A look back: SN 2000

Overview

What tools are available to study dynamic systems of many interacting agents?

Benchmark theory in economics: *dynamic games.*

But dynamic games can be hard to work with...

This talk is an example of the use of *mean field approximations* to simplify analysis of dynamic games.

Multiarmed bandit games

In this talk, we focus on *multiarmed bandit games.*

These are games where each agent faces a multiarmed bandit problem, but with rewards affected by *other agents'* actions.

We discuss a mean field approach to obtaining insight into equilibria of such games.

Outline

- Review: multiarmed bandits
- Multiarmed bandit games
- A mean field model
- Results:
  - Existence, uniqueness, convergence, approximation
- Related work

Review: Multiarmed bandits

Multiarmed bandits

Multiarmed bandits (MABs) are a canonical model for studying *learning in uncertain environments.*

Basic (stationary, stochastic) model:

At each time t, a single agent chooses one among n alternatives ("arms").

Alternative i returns a Bernoulli(θ_i) reward (i.i.d. across time and arms), where θ is *unknown*.

The objective is generally to learn the best arm "quickly".

Examples

Wireless channel selection:

Devices can choose one of n channels for transmission; channel quality is uncertain.

Product selection:

A firm can choose one of n products to sell or recommend in each period.

Online service selection:

An individual experiments with different online services each period (e.g., online gaming).

Optimal policies: examples

(1) Discounted expected reward criterion:

Assume agent discounts future rewards by $\beta < 1$.

Also assume the agent has a *prior* over θ.

Goal is to maximize $E[\sum_{t \geq 0} \beta^t \text{ Reward}_t \mid \text{prior}]$.

For this model, the *Gittins index policy* is optimal.

Optimal policies: examples

(2) Expected regret criterion:

Let θ_i* = max$_j$ θ_j.

Goal is to minimize:

E[Regret$_t$] = $t\theta_i$* - $\sum_{s \leq t}$E[Reward$_s$]

It is well known that optimal policies achieve

E[Regret$_t$] = O(log t) (e.g., Lai-Robbins, UCB).

State: aggregating history

Important observation:

Under the i.i.d. stationary reward model,
a sufficient statistic of the past history is:

$z_t = (w_t(1), \ell_t(1), \ldots, w_t(n), \ell_t(n))$

Where $w_t(i)$ = # of successes on arm i up to time t,
and $\ell_t(i)$ = # of failures on arm i up to time t.

Multiarmed bandit games

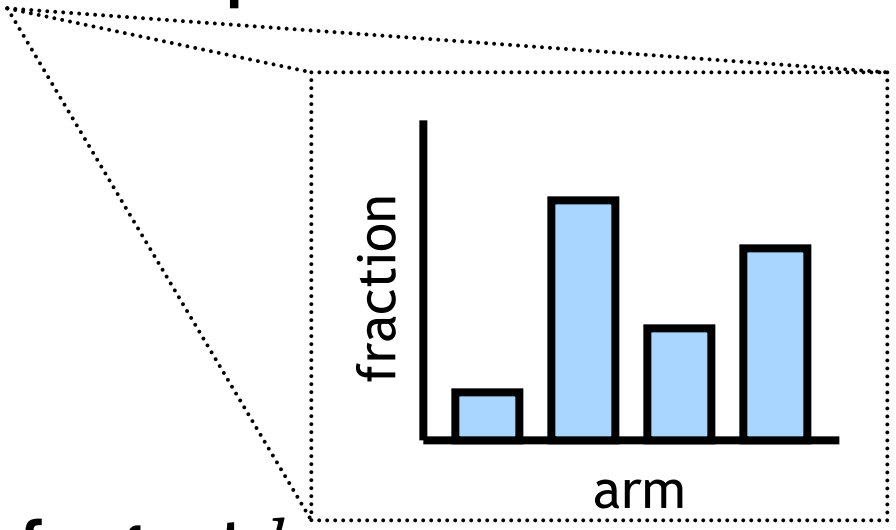
Multiple agents: MAB games

Now suppose m agents each play a multiarmed bandit, *but their rewards are coupled.*

Formally:

Let $f_t(i)$ = *fraction* of agents that pull arm i.

Agent k's reward on pulling arm i is Bernoulli(Q(θ_i^k, $f_t(i)$), independent across arms and time.



We call $\theta^k \in [0,1]^n$ the *type* of agent k .

Multiple agents: MAB games

This defines a *multiarmed bandit (MAB) game:*

Each individual has plays a multiarmed bandit, but with rewards affected by others.

All prior examples are really MAB games:

Wireless channel selection

Product selection

Online service selection

Examples of reward functions

Congestion models:

Q(θ, f) decreasing in f

e.g., wireless channel selection, product selection

Coordination models:

Q(θ, f) increasing in f

e.g., selection on online gaming service

Equilibrium: PBE

How should an agent play?

Observe that rewards are no longer stationary.

Dynamic game theory suggests that the right solution concept is *perfect Bayesian equilibrium:*

(1) An agent maintains *beliefs* over all that is unknown (including other agents' beliefs); and

(2) Chooses an optimal strategy (for their objective), given strategies chosen by other players.

Equilibria in dynamic games

PBE is *implausible:*

PBE makes very strong rationality assumptions, i.e., that agents track and forecast their competitors.

PBE is *intractable:*

Even finding an optimal strategy is intractable due to state space complexity—let alone an equilibrium.

An alternate approach

In "practice", such complex strategies are never implemented.

We consider an alternate question:

What happens if agents pretend the world is stationary, and play "simpler" strategies?

We use mean field approximations to provide insight into this question.

A mean field model

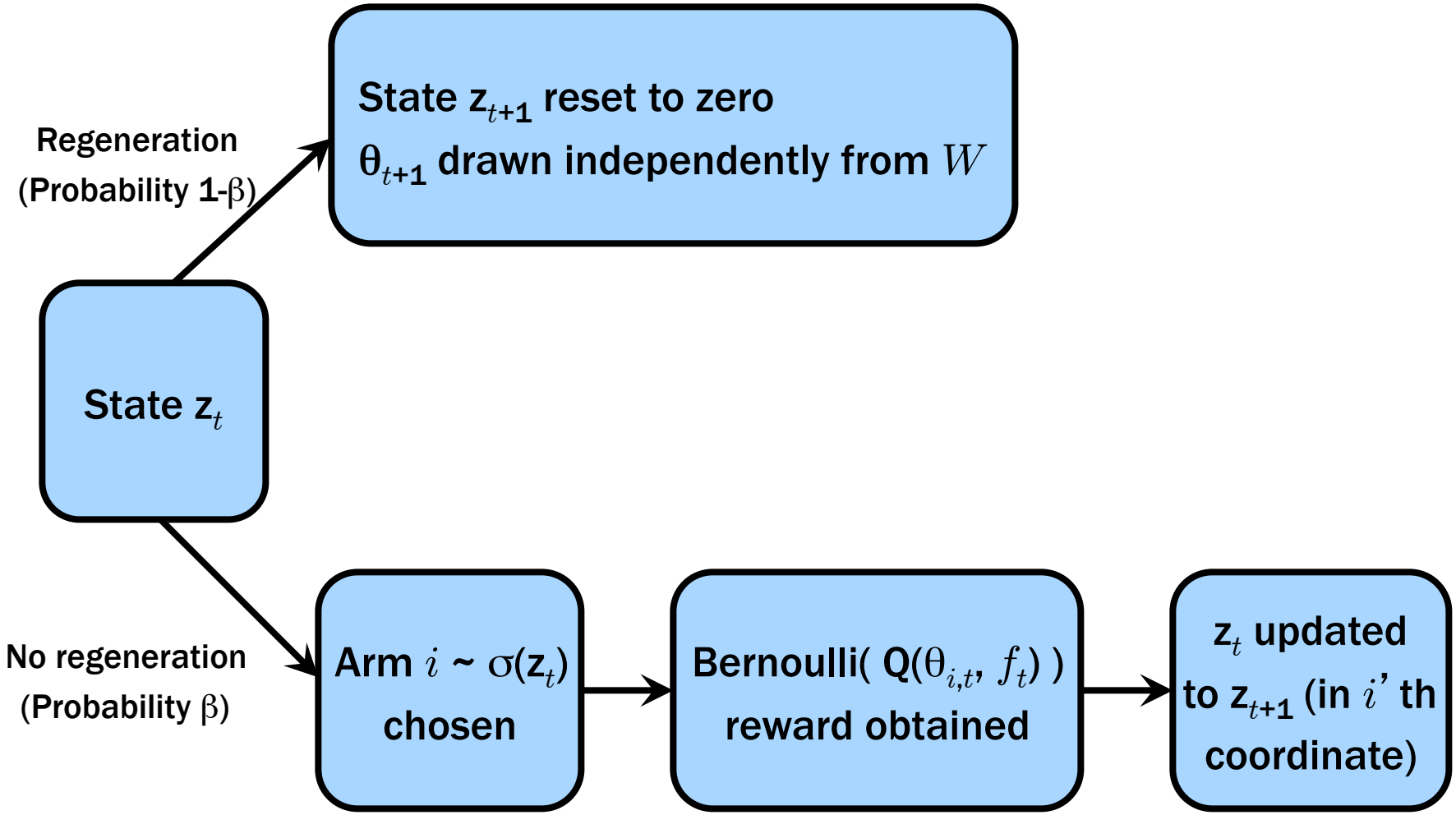
Mean field approach

We study the MAB game in a *mean field* model.

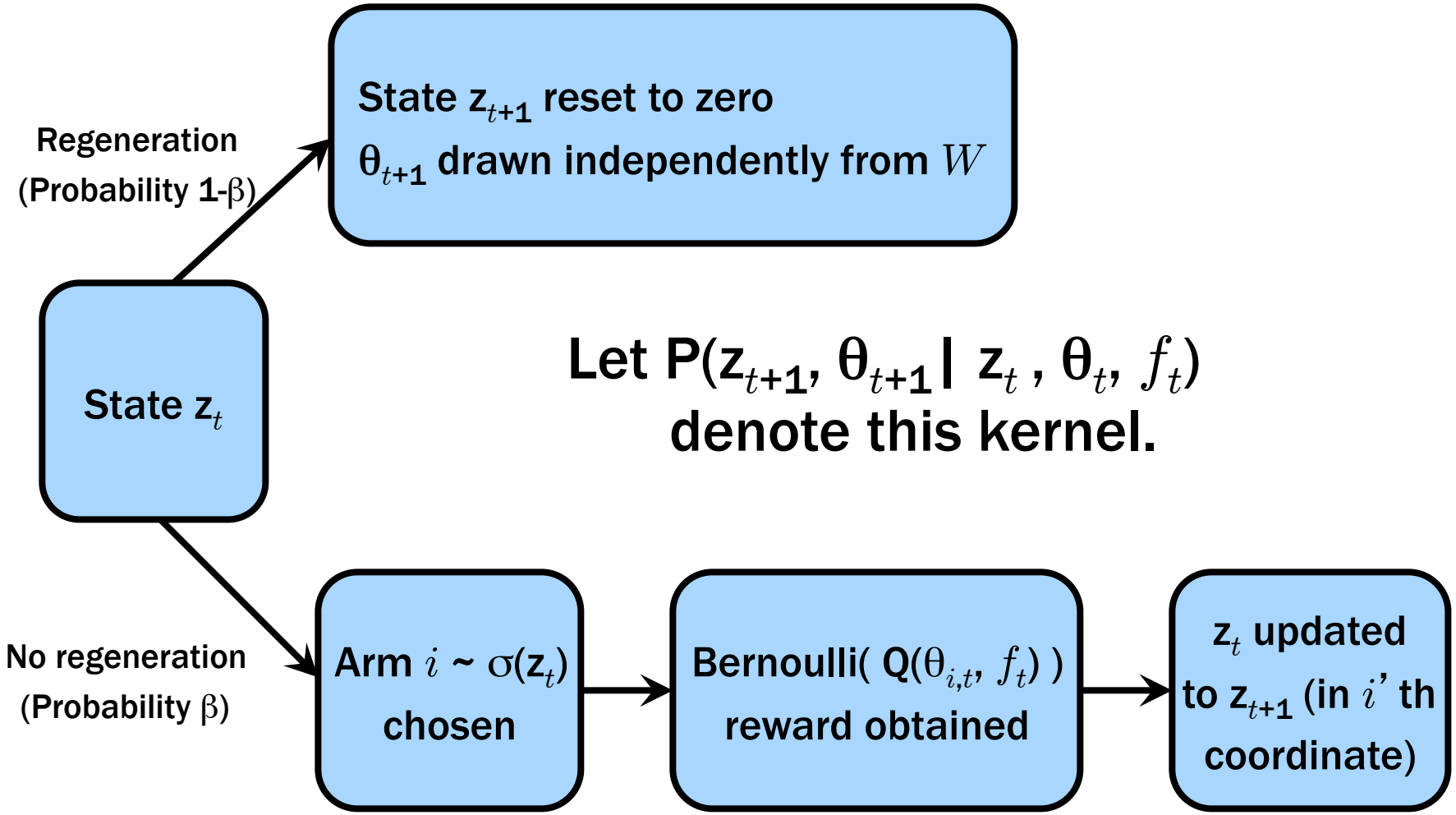
- An agent is characterized by state z_t and type θ_t.
- Agents "regenerate" after geometric(1-β) time.
  - Upon regeneration, θ sampled i.i.d. from a dist. W.
  - Upon regeneration, state reset to zero vector.
- Policy σ maps state z_t to (randomized) arm choice.
- Let f_t denote population profile at time t.

Note: all agents use the same policy σ.

Agent dynamics



Agent dynamics



Mean field dynamics

The mean field model for a policy σ is characterized by a sequence of joint distributions μ_t, $t \geq 0$, over states z and types θ.

μ_t(z, A) denotes the measure of agents at state z and with type $\in$ A at time t.

Mean field dynamics

(1) Given μ_t,

the population profile f_t is:

$$f_t(i) = \sum_z \int_\theta \sigma(z)(i)\, \mu_t(z, d\theta)$$

i.e., compose the measure μ_t with the policy σ.

(2) Given μ_t and f_t,

μ_{t+1} is obtained from agent dynamics:

$$\mu_{t+1}(z, A) = \sum_{z'} \int_\theta P(z, A \mid z', \theta, f_t)\, \mu_t(z', d\theta)$$

Mean field equilibrium

The preceding discussion motivates the notion of *mean field equilibrium:*

Given a policy σ, a measure μ is a MFE if it is a *fixed point of the mean field dynamics* (with associated MFE population profile f).

Discussion of MFE

(1) Note that in an MFE, the population profile is *fixed, and remains stationary over time.*

So if the world is in an MFE,
each agent solves a stationary stochastic MAB!

Discussion of MFE

(2) Note that no notion of *optimality* is part of the definition, because we fixed the policy *a priori*.

This allows us to determine whether "simple" policies yield meaningful behavior in MAB games (UCB, index policies, etc.).

But we might also introduce optimality into the definition itself.

Discussion of MFE

(3) Note that MFE is distinct from the literature on asymptotic *learning in games.*

In particular, in our model agents live for finite time and are always learning in steady state.

The learning in games literature focuses asymptotic behavior of agents, and whether they converge to a solution of the static game.

Results

Existence

Proposition:

If Q is continuous in f, an MFE exists.

Proof: **Brouwer's fixed point theorem.**

A contraction condition

Theorem:

Suppose Q is Lipschitz in f***, with Lipschitz constant*** **L.**

Then if:

$$\beta(1 + L) < 1,$$

the map $\mu_t \to \mu_{t+1}$ ***is a*** **contraction (in TV distance).**

(Note that this is true for *any* policy σ**!)**

Uniqueness and convergence

Corollary:

If $\beta(1 + L) < 1$, *then there exists a unique MFE, and mean field dynamics converge to it from any initial condition*.

Contraction: proof technique

- The proof relies on coupling characterization of total variation distance.
- Suppose μ, μ' have TV distance d.
- Observe that resulting population profiles f, f' have TV distance at most d.
- Construct $(z,\theta) \sim \mu$, $(z',\theta') \sim \mu'$, such that:

$$P((z,\theta) \neq (z',\theta')) = d$$

Contraction: proof technique

- Couple transitions and use Lipschitz condition
- Can show that TV distance at next time step is less than or equal to $\beta(d + (1\text{-}d)d\ L) \leq \beta(1+L)\ d$
- Why?

  Can couple so states/types at next time step differ only if:

  -no regeneration; and

  -initial states/types differed, or

  -initial states/types the same, and subsequent states differ

Contraction: discussion

Note that the condition is strong:

It requires either that

(1) agents do not live too long; or

(2) agents are not too sensitive to each other.

But this happens because the result applies for *any* policy.

Mean field limit

Theorem:

Let $\mu_t^{(m)}$ ***be the sequence of (random) joint state-type distributions in a system with*** m ***players, all using policy*** σ**.**

Under the same contraction condition (β **(1+L) < 1),** $\mu_t^{(m)}$ ***converges weakly to*** μ_t ***uniformly in*** t **(in L1).**

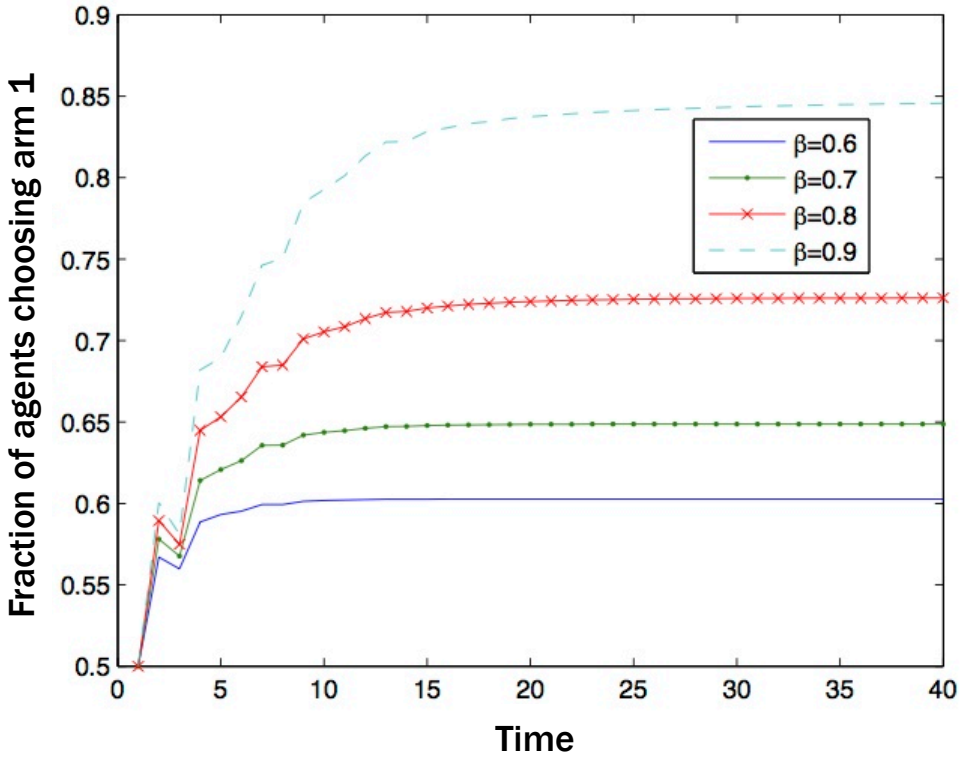
Intuition: **the contraction condition prevents the system from "drifting". (See also Glynn ' 04.)**

Numerics

The contraction condition appears to be very loose, based on numerical experiments.

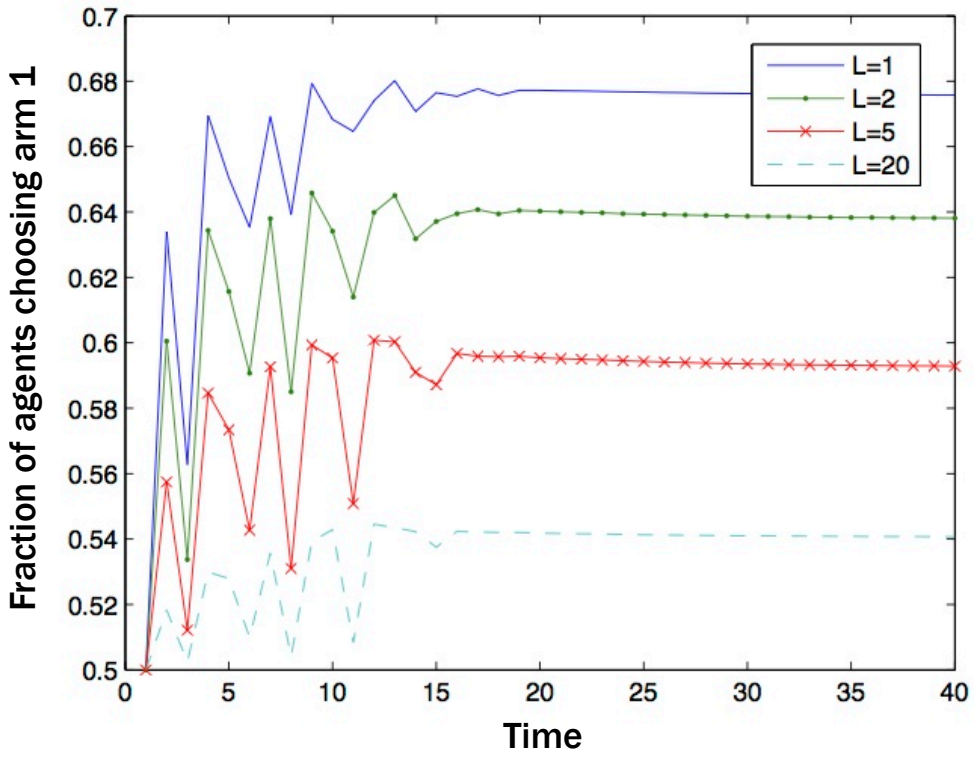
Numerics: large β

- $Q(\theta, f) = \theta f$, two arms, $E[\theta_1] = 0.8$, $E[\theta_2] = 0.33$



Numerics: large L

- $Q(\theta, f) = \theta/(1+Lf)$, two arms, $\beta = 0.9$, $E[\theta_1] = 0.8$, $E[\theta_2] = 0.33$



Decreasing rewards

As partial evidence of the conservatism of our result, we have recently established that:

(1) if σ is *positively sensitive* to rewards (informally, arms with higher rewards are more likely to be pulled); and
(2) if $Q(\theta, f)$ is decreasing in f,

then the MFE is unique.

We conjecture that the dynamics converge as well.

Conclusion

Summary

We are trying to study what happens when individuals learn *as if* the world is stationary, while in fact interactions with others create nonstationarity.

Our work shows that (under some conditions), the system eventually *becomes* stationary.

Mean field equilibria in games

More generally, our work illustrates the value of mean field equilibria in games:

Asymptotics vastly simplify the study of dynamic interactions among agents.

As a result, we can gain insight into previously intractable settings.

This makes mean field approximations to games invaluable tools for engineering of economic systems.

Coda: Related work

Mean field models arise in a variety of fields:

physics, applied math, engineering, economics…

Mean field models in dynamic games:

- *Economics:* Jovanovic and Rosenthal (1988); Stokey, Lucas, Prescott (1989); Hopenhayn (1992); Sleet (2002); Weintraub, Benkard, Van Roy (2008, 2010); Acemoglu and Jepsen (2010); Bodoh-Creed (2011)
- *Dynamic markets*: Wolinsky (1988); McAfee (1993); Backus and Lewis (2010); Iyer, Johari, Sundararajan (2011); Gummadi, Proutiere, Key (2012); Bodoh-Creed (2012); Duffie, Malamud, Manso (2009, 2010)
- *Control:* Glynn, Holliday, Goldsmith (2004); Lasry and Lions (2007); Huang, Caines, Malhame (2007-2012); Gueant (2009); Tembine, Altman, El Azouzi, le Boudec (2009); Yin, Mehta, Meyn, Shanbhag (2009); Adlakha, Johari, Weintraub (2009, 2011)

(Other names for MFE: Stationary equilibrium, oblivious equilibrium)