LEARNING DYNAMICS IN CONTINUOUS-TIME MEAN FIELD GAMES WITH CONGESTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study stationary equilibria and learning dynamics in a finite-state continuous-time mean field game with population-dependent action-completion rates generated by a smooth convex congestion potential. In this game, stationary equilibrium populations correspond one-to-one with optimizers of an Eisenberg–Gale-type convex program. We show that this variational structure also yields global convergence of the state–action dynamics under any fixed unichain policy, even though these dynamics are governed by a nonlinear Markov process. We then formulate an equivalent population game over deterministic unichain policies, where a concave ordinal potential gives convergence of best-response dynamics based on stationary long-run payoffs to the policy-equilibrium set. Finally, we study state–policy dynamics that track agents' physical states and committed policies under two revision mechanisms: exogenous opportunities arriving at a Poisson rate, and event-driven opportunities arising at action completions with a specified probability. Under both mechanisms, policy marginals converge to the policy-equilibrium set for every fixed admissible positive revision parameter. As the revision rate or probability tends to zero, the state–action populations associated with the omega-limit points of these dynamics approach the stationary mean field equilibrium populations. These results reveal how equilibrium emerges through repeated policy revisions and give our notion of stationary mean field equilibrium a dynamic foundation.

1 INTRODUCTION

Mean field games (MFGs) provide a powerful framework for studying learning and decision-making in large multi-agent systems (Yang et al., 2018; Zhang et al., 2025). They model strategic interactions through a representative agent's response to the aggregate population distribution (Huang et al., 2006; Lasry & Lions, 2007). This framework has been applied to driving and traffic routing (Huang et al., 2021), pedestrian crowd dynamics (Lachapelle & Wolfram, 2011), and cloud resource management (Mao et al., 2022). However, convergence of multi-agent learning is not guaranteed, even in the fundamental model-based setting studied here. Convergence guarantees still require additional structure, which can be difficult to identify in applications. In finite-horizon MFGs, representative analyses of fictitious play exploit potential structure (Cardaliaguet & Hadikhanloo, 2017) or monotonicity of population costs (Hadikhanloo & Silva, 2019; Perrin et al., 2020). For discounted MFGs with population-dependent transitions, learning guarantees have been established under strict contraction and regularity conditions (Guo et al., 2019; Xie et al., 2021; Zhang et al., 2025). Recent results give last-iterate convergence for finite-horizon games with population-independent transitions under Lasry–Lions monotonicity (Isobe et al., 2025) and single-loop learning for stationary games with herding structure (Zeng et al., 2025). In this work, we identify a congestion structure that is common in applications and supports both equilibrium computation and convergence of learning dynamics.

In transportation networks, ride-hailing platforms, charging systems, and service networks, agents often choose their next action when they complete a trip, charging session, or service task. These decision times differ across agents. Moreover, congestion changes these action durations: traffic delays trips (Roughgarden & Tardos, 2002), shared charging capacity affects charging times (Aveklouris et al., 2019), and competing jobs prolong service in processor-sharing queues (Anselmi et al., 2013). Thus congestion affects both state-transition speeds and how often agents have opportunities to revise their policies at action completions. Therefore, we study a finite-state continuous-time MFG with population-dependent completion rates, where each agent waits until its current action

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is completed before choosing the next one. We assume that these rates are generated by a smooth convex congestion potential that describes how congestion changes action durations. This choice is motivated by Kelly's throughput-based approximations for closed queueing networks (Kelly, 1989). It captures symmetric marginal congestion effects in the systems described above.

A central question here is how a properly defined equilibrium can emerge from asynchronous decentralized learning. We focus on stationary MFG equilibria, which require both individual optimality and population consistency: given the stationary population, the policy maximizes an agent's long-run average reward from every initial state, and the population is invariant under that policy. However, these optimality and consistency conditions form a nonlinear fixed point, making equilibrium computation difficult even with finite state and action spaces (Neumann, 2020). Establishing convergence to equilibrium through decentralized learning is harder still: even under a fixed policy, population-dependent completion rates produce a nonlinear Markov process, for which a finite state space alone does not ensure convergence (Neumann, 2023). Constructing a Lyapunov function adapted to this feedback is a central difficulty; available constructions and local-stability results rely on additional structure (Budhiraja et al., 2015a;b). When agents are allowed to change policies, the population may still be adjusting to earlier choices when further updates occur. Proving convergence then requires analyzing policy updates and population dynamics together. A natural asynchronous learning rule lets agents update their policies when they complete an action. This rule adds another layer of complexity: congestion affects both how quickly agents act and how often they can update their policies. We address these challenges through the common congestion potential, drawing on a line of research on equilibrium characterization and learning dynamics in related models. We develop the Eisenberg–Gale formulation and policy-space learning approach of Courcoubetis & Dimakis (2025) for continuous-time Markov dynamics with a general smooth convex congestion potential in place of their linear execution-cost term. Our state–policy formulation also appears in Pedroso et al. (2025a;b), who use a joint description of agents' states and policies; our event-driven revision is inspired by the myopic adjustment process of Neumann (2024). Building on these works, we identify a class of MFGs with a common congestion potential and establish convergence of their learning dynamics. Our contributions are summarized as follows.

Contributions. (i) We formulate the congestion-dependent continuous-time MFG and prove a one-to-one correspondence between stationary MFG equilibrium populations and the optimal solutions of a finite-dimensional Eisenberg–Gale-type convex program. This correspondence yields equilibrium existence and a direct computational method. (ii) We first study state–action dynamics under any fixed stationary mixed unichain policy. Using the congestion potential, we construct a Lyapunov function and prove global convergence to the unique stationary population. (iii) We then consider policy revision based on stationary payoffs and formulate an equivalent population game over deterministic unichain policies. Its concave ordinal potential gives global best-response convergence to the policy-equilibrium set, whose associated stationary populations are exactly the stationary MFG equilibrium populations. (iv) Finally, we study state–policy dynamics with exogenous revision and event-driven state–policy dynamics, with revision opportunities arriving at a Poisson rate or arising at action completions with a specified probability, respectively. Under either mechanism, policy marginals converge to the policy-equilibrium set for each fixed admissible positive revision parameter. As the revision rate or probability tends to zero, the state–action populations associated with the omega-limit points of these dynamics approach the stationary MFG equilibrium populations.

2 Continuous-Time MFG and Stationary Equilibria

We consider an MFG where each agent faces a continuous-time Markov decision process (MDP). The agent selects an action and receives its next opportunity to choose an action when the current one is completed. Agents interact through action completion rates, which depend on the current state–action population. Let $\mathcal{S}$ and $\mathcal{A}$ be finite nonempty state and action spaces, $\mathcal{I} := \mathcal{S} \times \mathcal{A}$, and let the population mass be $m > 0$. Write $\mu \in \mathcal{M}_m(\mathcal{I})$ for the state–action population.[1] For $i = (s, a)$, μ_i is the mass currently holding action a in state s. An agent holding pair i completes its action at instantaneous rate $\lambda_i(\mu) > 0$. For a fixed population μ, its action duration is exponentially distributed with rate $\lambda_i(\mu)$. The state–action completion rate is $x_i(\mu) = \mu_i \lambda_i(\mu)$, measured in mass per unit

[1] For any finite set E, write $\mathcal{P}(E)$ for its probability simplex and $\mathcal{M}_m(E) := \{\mu \in \mathbb{R}_+^E : \sum_{e \in E} \mu_e = m\} = m\mathcal{P}(E)$ for its mass-m simplex.

time. Completion yields reward $r_{sa} > 0$ and a transition to s' with probability $P^a_{ss'}$, where P^a is a transition matrix on $\mathcal{S}$ for each $a \in \mathcal{A}$. We specify the congestion dependence of action completion rates through a convex potential G.

Assumption 1 (Potential congestion)**.** *There is an open convex neighborhood $U \supset \mathbb{R}^{\mathcal{I}}_+$ and a convex $G \in C^2(U)$ such that $\tau_i(x) = \partial_i G(x) > 0$ for every $x \in \mathbb{R}^{\mathcal{I}}_+$ and $i \in \mathcal{I}$.*

We call $\tau_i(x)$ the sojourn-time parameter in pair i, or simply its sojourn time, and set $\lambda_i(\mu) = 1/\tau_i(x(\mu))$. The potential implies symmetric marginal effects, $\partial_j \tau_i = \partial_i \tau_j$, while convexity gives $\nabla^2 G \succeq 0$. Combining the rate specification with $x_i(\mu) = \mu_i \lambda_i(\mu)$ yields

$$\mu_i = x_i \tau_i(x) = x_i \partial_i G(x). \tag{1}$$

Thus the state–action completion rates $x_i(\mu)$ and individual completion rates $\lambda_i(\mu)$ at population μ are determined by solving this implicit relation for x. The potential provides a variational solution: for $\mu \in \mathbb{R}^{\mathcal{I}}_+$, define[2]

$$x(\mu) := \arg\max_{x \in \mathbb{R}^{\mathcal{I}}_+} \sum_{i \in \mathcal{I}} \mu_i \log x_i - G(x). \tag{2}$$

Proposition 2. *Under Assumption 1,* (2) *has a unique optimizer, which is also the unique nonnegative solution of* (1)*. At every $\mu \in \mathbb{R}^{\mathcal{I}}_+$, the map $\mu \mapsto x(\mu)$ is the restriction of a C^1 map defined on an open neighborhood of μ. Consequently, $\lambda_i(\mu) := 1/\partial_i G(x(\mu))$ satisfies $x_i(\mu) = \mu_i \lambda_i(\mu)$. Both x and λ are locally Lipschitz on $\mathbb{R}^{\mathcal{I}}_+$ and globally Lipschitz on $\mathcal{M}_m(\mathcal{I})$.*

The proposition ensures that the completion rates are uniquely defined and regular even when some state–action pairs have zero mass. Fix a continuous population trajectory $\mu : [0, \infty) \to \mathcal{M}_m(\mathcal{I})$, which an individual agent takes as given. A time-dependent Markov policy $\pi_t(s, \cdot) \in \mathcal{P}(\mathcal{A})$, measurable in t, gives the action probabilities for an agent choosing its next action in state s at time t. The selected action is held until completion. The agent's state–action pair is then a time-inhomogeneous Markov chain with generator $Q^\pi(\mu(t), t)$, where

$$Q^\pi_{ij}(\mu, t) = \lambda_i(\mu)\big(P^a_{ss'}\pi_t(s', a') - \delta_{ij}\big), \qquad i = (s, a), \quad j = (s', a').$$

Here δ_{ij} is the Kronecker delta. The off-diagonal entries are the nonnegative transition rates between distinct state–action pairs, and each row sums to zero. At each completion, the agent moves according to P^a and selects its next action using the policy at the arrival state and time. Completions that return to the same state–action pair are included in the completion rate $\lambda_i(\mu)$ and still generate reward. For a stationary policy, $\pi_t(s, a) = \pi(s, a)$, and we write $Q^\pi(\mu)$ without the time argument.

More generally, an admissible policy $\tilde{\pi}$ may randomize and use the decision time and the history observed so far, but must hold each selected action until completion.[3] Let $N(T)$ count the action completions of an individual agent by time T, and let (S_{k-1}, A_k) be the state–action pair it holds before the kth completion. The expected long-run reward rate of an agent starting in state s is

$$J^{\tilde{\pi}}(s; \mu(\cdot)) := \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}^{\tilde{\pi}, \mu(\cdot)}_s \sum_{k=1}^{N(T)} r_{S_{k-1}A_k}.$$

The expectation is over the agent's completion times, state transitions, and action choices. The agent maximizes this reward rate over admissible policies, taking the entire population trajectory as given. For a constant population trajectory $\mu(t) \equiv \mu$, write $J^{\tilde{\pi}}(s; \mu)$ for the same criterion. In this case, the sojourn times $\tau_i = \tau_i(x(\mu))$ and completion rates $1/\tau_i$ are fixed. The *individual optimization problem* is the agent's finite continuous-time Markov decision process with these fixed rates and the long-run average-reward criterion $J^{\tilde{\pi}}(s; \mu)$, measured per unit time.

Definition 3 (Stationary MFG equilibrium)**.** A pair $(\mu^*, \pi^*) \in \mathcal{M}_m(\mathcal{I}) \times \mathcal{P}(\mathcal{A})^{\mathcal{S}}$ is a *stationary MFG equilibrium* if the following conditions hold: (i) *Individual optimality.* For every initial state $s \in \mathcal{S}$ and every admissible policy $\tilde{\pi}$, $J^{\pi^*}(s; \mu^*) \geq J^{\tilde{\pi}}(s; \mu^*)$. (ii) *Population consistency.* The state–action population μ^* is invariant under the generator $Q^{\pi^*}(\mu^*)$: $\sum_{i \in \mathcal{I}} \mu^*_i Q^{\pi^*}_{ij}(\mu^*) = 0$ for every $j \in \mathcal{I}$. We denote the set of stationary MFG equilibrium pairs by $\mathcal{E}^*$.

[2] We use the convention $0 \log z := 0$ for $z \geq 0$ and $c \log 0 := -\infty$ for $c > 0$. Hence $\operatorname{supp} x(\mu) = \operatorname{supp} \mu$: positive μ_i excludes $x_i = 0$, while for $\mu_i = 0$ any positive x_i lowers the objective because $\partial_i G > 0$.

[3] Formally, the action-selection probabilities are Borel-measurable functions of the observed states, actions, and decision times. This condition ensures that the resulting process and expected reward are well defined.

Condition (i) means that no agent can improve its long-run reward rate by deviating from π^* when the population is fixed at μ^*, regardless of its initial state. Condition (ii) means that, if all agents follow π^*, a population initialized at μ^* remains unchanged. We impose the standard communication condition that makes the optimal long-run reward rate in this MDP with fixed sojourn times independent of the initial state.

Assumption 4 (Weak communication)**.** *The state space admits a partition $\mathcal{S} = \mathcal{S}_0 \sqcup \mathcal{S}_1$ such that every state in $\mathcal{S}_0$ is transient under every stationary policy, while for any $s, s' \in \mathcal{S}_1$ there exists a deterministic stationary policy under which s' is accessible from s.*

For fixed sojourn times, weak communication guarantees an optimal policy for the individual optimization problem. The optimal policy can be found by solving a linear program (see Appendix A.2). The potential makes the coupled conditions (i) and (ii) in Definition 3 computationally tractable: state balance and individual optimality are jointly encoded by the Karush–Kuhn–Tucker (KKT) conditions of

$$\max_{x \geq 0,\ r \cdot x > 0} \ m \log(r \cdot x) - G(x) \quad \text{s.t.} \quad \sum_{s',a} x_{s'a}(\delta_{ss'} - P^a_{s's}) = 0, \quad s \in \mathcal{S}. \tag{3}$$

Let $\mathcal{X}^*$ denote the set of optimal solutions to (3), and let $\mathcal{E}^*_{\mathcal{I}}$ denote the set of state–action populations μ for which there is a stationary policy π with $(\mu, \pi) \in \mathcal{E}^*$.

Theorem 5 (Eisenberg–Gale characterization of stationary equilibria)**.** *Under Assumptions 1 and 4, the following statements hold.* (i) *Program* (3) *has an optimal solution. The map*[4] *$x \mapsto x \odot \nabla G(x)$ is a bijection from $\mathcal{X}^*$ to $\mathcal{E}^*_{\mathcal{I}}$, with inverse $\mu \mapsto x(\mu)$. For each $x^* \in \mathcal{X}^*$, set $\pi^*(s,a) = x^*_{sa} / \sum_{a'} x^*_{sa'}$ at every state with $\sum_{a'} x^*_{sa'} > 0$, and $\pi^*(s,a) = 1/|\mathcal{A}|$ at every remaining state. Then $(x^* \odot \nabla G(x^*), \pi^*)$ is a stationary MFG equilibrium.* (ii) *There are $g^* > 0$ and $\tau^* \in (0,\infty)^{\mathcal{I}}$ such that $r \cdot x^* = m g^*$ and $\nabla G(x^*) = \tau^*$ for every $x^* \in \mathcal{X}^*$. For every stationary MFG equilibrium (μ^*, π^*), the average reward satisfies $J^{\pi^*}(s; \mu^*) = g^*$ for every initial state $s \in \mathcal{S}$, and the individual completion rates satisfy $\lambda_i(\mu^*) = 1/\tau^*_i$ for every $i \in \mathcal{I}$.* (iii) *The sets $\mathcal{X}^*$ and $\mathcal{E}^*_{\mathcal{I}}$ are nonempty, compact, and convex. More precisely, $\mathcal{E}^*_{\mathcal{I}} = \{\tau^* \odot x : x \in \mathcal{X}^*\}$.*

The theorem reduces equilibrium computation to solving a convex optimization problem. Although the equilibrium population need not be unique, all equilibria give agents the same long-run reward rate and the same individual completion rate at each state–action pair. Moreover, any convex combination of equilibrium population vectors is also an equilibrium population. Having characterized the equilibrium target, we now ask how it can emerge through state–action dynamics and policy revision.

3 State–Action Dynamics under a Fixed Policy

We first study how the population adjusts when all agents use the same stationary policy. Because the completion rates depend on congestion, the population follows a nonlinear Markov process, whose convergence is not guaranteed in general. Fix a stationary mixed unichain policy π.[5] The corresponding embedded state–action chain has state space $\mathcal{I}^\pi = \{(s,a) : \pi(s,a) > 0\}$ and transition matrix $p^\pi_{ij} := P^a_{ss'}\pi(s',a')$, where $i = (s,a)$ and $j = (s',a')$. Let $\mathcal{S}^\pi_{\mathrm{rec}} \subseteq \mathcal{S}$ be the recurrent state class and $\mathcal{S}^\pi_{\mathrm{tr}} := \mathcal{S} \setminus \mathcal{S}^\pi_{\mathrm{rec}}$ the transient states. The corresponding state–action sets are $\mathcal{I}^\pi_{\mathrm{rec}} := \{(s,a) \in \mathcal{I}^\pi : s \in \mathcal{S}^\pi_{\mathrm{rec}}\}$ and $\mathcal{I}^\pi_{\mathrm{tr}} := \mathcal{I}^\pi \setminus \mathcal{I}^\pi_{\mathrm{rec}}$. Under p^π, $\mathcal{I}^\pi_{\mathrm{rec}}$ is the unique recurrent class: an agent reaches it with probability one and cannot leave, while pairs in $\mathcal{I}^\pi_{\mathrm{tr}}$ are visited only transiently. We consider state–action populations compatible with π, meaning that no mass is assigned to actions that the policy never selects: $\mathcal{M}^\pi_m := \{\mu \in \mathcal{M}_m(\mathcal{I}) : \mu_i = 0 \text{ for } i \notin \mathcal{I}^\pi\}$. The state–action population under π evolves according to the nonlinear Kolmogorov forward equation

$$\dot{\mu}_j = \sum_{i \in \mathcal{I}^\pi} x_i(\mu) p^\pi_{ij} - x_j(\mu), \qquad j \in \mathcal{I}^\pi. \tag{4}$$

At each pair j, the population changes by the arrival rate $\sum_i x_i(\mu) p^\pi_{ij}$ minus the completion rate $x_j(\mu)$. The congestion potential determines these rates through (2), and therefore governs how quickly the population evolves.

[4] Here $\odot$ denotes componentwise multiplication: $(u \odot v)_i = u_i v_i$.

[5] The embedded state chain records the state reached after each completed action; its transition matrix is $P^\pi_{ss'} = \sum_a \pi(s,a) P^a_{ss'}$. A policy is unichain if its embedded state chain has exactly one recurrent class, possibly with transient states. Within the recurrent class, all states communicate and the chain cannot leave; states outside it are visited only finitely often almost surely. See Puterman (1994, Ch. 8).

The unique invariant state probability distribution $q^\pi \in \mathcal{P}(\mathcal{S})$ is positive on $\mathcal{S}^\pi_{\mathrm{rec}}$ and zero on $\mathcal{S}^\pi_{\mathrm{tr}}$. Define the *invariant state–action probability distribution* of the embedded chain by $\rho^\pi_{sa} := q^\pi_s \pi(s,a)$ for all $(s,a) \in \mathcal{I}$. Here ρ^π_{sa} is the long-run fraction of visits to (s,a) in that chain, which takes one step per completed action. Then $\rho^\pi \in \mathcal{P}(\mathcal{I})$, and its restriction to $\mathcal{I}^\pi$ is the unique invariant probability distribution of the state–action transition matrix p^π. Under Assumption 1, there is a unique stationary total completion rate $\bar{\Lambda}^\pi > 0$ such that $\bar{x}^\pi := \bar{\Lambda}^\pi \rho^\pi$ satisfies $\bar{x}^\pi \cdot \nabla G(\bar{x}^\pi) = m$, and $\bar{\mu}^\pi := \bar{x}^\pi \odot \nabla G(\bar{x}^\pi)$ is stationary for (4); see Lemma 21 in Appendix B.3. The following proposition constructs a Lyapunov function for each stationary unichain policy π.

Proposition 6. *Under Assumption 1, let $\bar{x}^\pi$ and $\bar{\mu}^\pi$ be as above for the fixed unichain policy π. (i) Let $f^*(\mu)$ be the optimal value in* (2)*. For $\mu \in \mathcal{M}^\pi_m$ with $\mu_i = 0$ for every $i \in \mathcal{I}^\pi_{\mathrm{tr}}$, define*

$$L_\pi(\mu) := f^*(\mu) - \sum_{i \in \mathcal{I}^\pi_{\mathrm{rec}}} \mu_i \log \bar{x}^\pi_i + G(\bar{x}^\pi). \tag{5}$$

Then $L_\pi(\mu) \geq 0$, with equality if and only if $\mu = \bar{\mu}^\pi$. (ii) Along every solution of (4) *with $\mu_i > 0$ for all $i \in \mathcal{I}^\pi_{\mathrm{rec}}$ and $\mu_i = 0$ for all $i \in \mathcal{I}^\pi_{\mathrm{tr}}$, write $x = x(\mu(t))$. Then $\frac{d}{dt} L_\pi(\mu(t)) = \sum_{i,j \in \mathcal{I}^\pi_{\mathrm{rec}}} x_i p^\pi_{ij} \log \frac{x_j/\bar{x}^\pi_j}{x_i/\bar{x}^\pi_i} \leq 0$. The derivative is zero if and only if $\mu(t) = \bar{\mu}^\pi$.*

If some agents initially occupy transient pairs, the total mass remaining on those pairs decays exponentially (Lemma 19 in Appendix B). Combining this decay with the Lyapunov argument yields the following global convergence theorem.

Theorem 7 (Global convergence under a fixed unichain policy)**.** *Under Assumption 1, fix any stationary mixed unichain policy π. For every initial condition in $\mathcal{M}^\pi_m$, including its boundary, the nonlinear Kolmogorov forward equation* (4) *has a unique global solution. The equation has exactly one stationary population $\bar{\mu}^\pi$, and every solution satisfies $\mu(t) \to \bar{\mu}^\pi$ and $x(\mu(t)) \to \bar{x}^\pi$.*

4 Policy Dynamics in an Equivalent Population Game

We study policy revision assuming that the state–action population adjusts instantaneously to the stationary population associated with the current policy masses. This defines a population game whose strategies are policies and whose payoffs are long-run reward rates. We use deterministic unichain policies as strategies, so each has a unique invariant state probability distribution:

$$\Pi := \{\pi : \mathcal{S} \to \mathcal{A} : \ P^\pi \text{ has exactly one recurrent class}\}, \tag{6}$$

where $P^\pi_{ss'} = P^{\pi(s)}_{ss'}$. Under weak communication, Π contains an optimal policy for each fixed positive vector of sojourn times τ (Lemma 22 in Appendix C). A policy distribution $\nu \in \mathcal{M}_m(\Pi)$ assigns mass ν_π to policy π. Using the invariant state probability distribution q^π from Section 3, write $\rho^\pi_{sa} = q^\pi_s \mathbf{1}_{\{a=\pi(s)\}}$ and $r^\pi := r \cdot \rho^\pi$ for its invariant state–action probability distribution and expected reward per completion. Let $\mathbf{R} \in \mathbb{R}^{\mathcal{I} \times \Pi}$ have columns ρ^π. For policy completion rates $\chi \in \mathbb{R}^\Pi_+$, the state–action completion rates are $\mathbf{R}\chi$, and policy π's mean time per completion is $\tau^\pi(\chi) := \sum_i \rho^\pi_i \tau_i(\mathbf{R}\chi)$. The stationary rates $\bar{\chi}^\nu$ are the unique solution of the mass-balance equations $\nu_\pi = \bar{\chi}^\nu_\pi \tau^\pi(\bar{\chi}^\nu)$, or, equivalently, the optimizer

$$\bar{\chi}^\nu := \arg\max_{\chi \in \mathbb{R}^\Pi_+} \left\{ \sum_{\pi \in \Pi} \nu_\pi \log \chi_\pi - G(\mathbf{R}\chi) \right\}, \tag{7}$$

using the extended-value convention in (2). These rates determine the stationary state–action completion rates and population

$$\bar{x}^\nu := \mathbf{R}\bar{\chi}^\nu, \qquad \bar{\mu}^\nu := \bar{x}^\nu \odot \nabla G(\bar{x}^\nu). \tag{8}$$

The mass balances give $\bar{\mu}^\nu \in \mathcal{M}_m(\mathcal{I})$. The payoff of strategy π under the policy distribution ν is its long-run reward rate in this stationary environment, $H_\pi(\nu) := r^\pi / \tau^\pi(\bar{\chi}^\nu) = r^\pi / \big(\sum_i \rho^\pi_i \tau_i(\bar{x}^\nu)\big) > 0$. Although $\bar{\chi}^\nu_\pi = 0$ exactly when $\nu_\pi = 0$, the payoff is defined for every policy, including unused ones. Write $H(\nu) := (H_\pi(\nu))_{\pi \in \Pi}$ for the payoff vector.

We assume that agents know the game structure and can evaluate the current policy distribution ν and its associated stationary state–action population $\bar{\mu}^\nu$. Agents receive unit-rate opportunities to

switch to a best response. Define $\mathrm{BR}_\Pi(\nu) := \arg\max_{\pi\in\Pi} H_\pi(\nu)$, and the corresponding mass-m best-response correspondence and equilibrium set by

$$\Gamma(\nu) := \{\gamma \in \mathcal{M}_m(\Pi) : \mathrm{supp}(\gamma) \subseteq \mathrm{BR}_\Pi(\nu)\}, \qquad \mathcal{E}^*_\Pi := \{\nu : \nu \in \Gamma(\nu)\}. \tag{9}$$

At equilibrium, every policy with positive mass is a best response. The resulting policy dynamics are $\dot\nu \in \Gamma(\nu) - \nu$. The inclusion allows measurable choices among maximizing policies; convergence of every solution follows from the potential

$$\Phi(\nu) := \max_{\chi\geq 0} \left\{ \sum_{\pi:\nu_\pi>0} \nu_\pi \log\left(\frac{r^\pi \chi_\pi}{\nu_\pi}\right) - G(\mathbf{R}\chi) \right\}. \tag{10}$$

Write $\Phi^* := \max_{\nu\in\mathcal{M}_m(\Pi)} \Phi(\nu)$. Define the potential gap $\Delta(\nu) := \Phi^* - \Phi(\nu)$ and the best-response gap for logarithmic payoffs $M_{\mathrm{BR}}(\nu) := m\max_\pi \log H_\pi(\nu) - \sum_\pi \nu_\pi \log H_\pi(\nu)$.

Theorem 8 (Potential structure and convergence of policy dynamics)**.** *Under Assumptions 1 and 4, the following statements hold. (i) The map $\nu \mapsto \bar\chi^\nu$ admits local C^1 extensions, and Φ admits local C^2 extensions, at every point of $\mathcal{M}_m(\Pi)$, with $\partial\Phi(\nu)/\partial\nu_\pi = \log H_\pi(\nu) - 1$. (ii) The function Φ is a concave ordinal potential for the population game with payoff map H,[6] and $\mathcal{E}^*_\Pi = \arg\max_{\nu\in\mathcal{M}_m(\Pi)} \Phi(\nu)$ is nonempty, compact, and convex. (iii) The best-response inclusion $\dot\nu \in \Gamma(\nu) - \nu$ has a global absolutely continuous solution from every initial point of $\mathcal{M}_m(\Pi)$. Every such solution has an ω-limit set $\omega(\nu) \subseteq \mathcal{E}^*_\Pi$.[7] In particular, along each solution, $\dot\Phi = M_{\mathrm{BR}} \geq \Delta \geq 0$ almost everywhere, $\Delta(\nu(t)) \leq e^{-t}\Delta(\nu(0))$, and $M_{\mathrm{BR}}(\nu(t)) \leq e^{-t}M_{\mathrm{BR}}(\nu(0))$ for all $t \geq 0$. The set $\mathcal{E}^*_\Pi$ is globally asymptotically stable: it is Lyapunov stable and attracts every solution.*

We next identify the stationary MFG populations associated with policy equilibria in $\mathcal{E}^*_\Pi$. Since agents choose actions at completions, a stationary mixed policy is recovered by normalizing the completion rates:

$$\bar\pi^\nu(s,a) := \frac{\bar x^\nu_{sa}}{\sum_{a'} \bar x^\nu_{sa'}}, \qquad \sum_{a'} \bar x^\nu_{sa'} > 0. \tag{11}$$

At states with zero completion rate, the uniform action choice from Theorem 5(i) completes the equilibrium policy.

Theorem 9 (Policy equilibria and stationary MFG equilibria)**.** *Under Assumptions 1 and 4, the following statements hold. (i) For every $\nu^* \in \mathcal{E}^*_\Pi$, the rule in* (11) *can be completed by choosing action probabilities $1/|\mathcal{A}|$ at states with $\sum_a \bar x^{\nu^*}_{sa} = 0$ to obtain a stationary mixed policy $\bar\pi^{\nu^*}$ with $(\bar\mu^{\nu^*}, \bar\pi^{\nu^*}) \in \mathcal{E}^*$. Conversely, every $(\mu^*, \pi^*) \in \mathcal{E}^*$ admits a $\nu^* \in \mathcal{E}^*_\Pi$ such that $\mu^* = \bar\mu^{\nu^*}$ and π^* agrees with* (11) *wherever $\sum_a \bar x^{\nu^*}_{sa} > 0$. Thus $\mathcal{E}^*_\mathcal{I} = \{\bar\mu^\nu : \nu \in \mathcal{E}^*_\Pi\}$, where $\mathcal{E}^*_\mathcal{I}$ is defined before Theorem 5. (ii) Let τ^*, g^* be as in Theorem 5, and set $\tau^{\pi,*} := \sum_i \rho^\pi_i \tau^*_i$, $H^*_\pi := r^\pi/\tau^{\pi,*}$, and $H^* := (H^*_\pi)_{\pi\in\Pi}$. For every $\nu^* \in \mathcal{E}^*_\Pi$ and $\pi \in \Pi$, $\bar\chi^{\nu^*}_\pi = \nu^*_\pi/\tau^{\pi,*}$ and $H_\pi(\nu^*) = H^*_\pi$, with $\max_\pi H^*_\pi = g^*$. (iii) For any trajectory $\nu : [0,\infty) \to \mathcal{M}_m(\Pi)$ satisfying $\mathrm{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$, $\mathrm{dist}(\bar\mu^{\nu(t)}, \mathcal{E}^*_\mathcal{I}) \to 0$, $H(\nu(t)) \to H^*$, $\lambda(\bar\mu^{\nu(t)}) \to (1/\tau^*_i)_{i\in\mathcal{I}}$, and $r\cdot\bar x^{\nu(t)}/m \to g^*$.*

Together, the two theorems show that best-response learning makes $\bar\mu^{\nu(t)}$ approach $\mathcal{E}^*_\mathcal{I}$. We next let the state–action population and policy distribution evolve together, allowing the population to lag behind its associated stationary population.

5 State–Policy Dynamics with Exogenous Revision

Agents in the same physical state can follow different policies, so describing the state–policy dynamics requires tracking their state and committed policy together. Let $\eta = (\eta_{s,\pi})_{s\in\mathcal{S},\pi\in\Pi} \in \mathcal{M}_m(\mathcal{S}\times\Pi)$ be the state–policy population, where $\eta_{s,\pi}$ is the mass in state s committed to policy π. It induces the policy marginal and state–action population

$$\nu_\pi(\eta) := \sum_s \eta_{s,\pi}, \qquad \mu_{sa}(\eta) := \sum_{\pi:\pi(s)=a} \eta_{s,\pi}. \tag{12}$$

[6] Here "ordinal" means that $\partial\Phi/\partial\nu_\pi - \partial\Phi/\partial\nu_{\pi'}$ has the same sign as $H_\pi - H_{\pi'}$: an infinitesimal shift of mass toward a policy with higher payoff increases Φ (see Monderer & Shapley, 1996, Section 2).

[7] For a trajectory $z(\cdot)$, the ω-limit set $\omega(z)$ consists of all points z^∞ such that $z(t_n) \to z^\infty$ along some sequence $t_n \to \infty$.

We omit the argument η when the current state–policy population is clear. The state–policy completion-rate vector $X(\eta) \in \mathbb{R}_+^{\mathcal{S}\times\Pi}$ has components

$$X_{s,\pi}(\eta) := \eta_{s,\pi}\lambda_{s,\pi(s)}(\mu). \tag{13}$$

Summing over policies that prescribe action a in state s gives $x_{sa}(\mu) = \sum_{\pi:\pi(s)=a} X_{s,\pi}(\eta)$.

We first give each agent exogenous revision opportunities at Poisson rate κ. At such an opportunity, it keeps its physical state and selects a policy $\pi \in \mathrm{supp}(\gamma)$ with probability γ_π/m, where $\gamma \in \Gamma(\nu)$. It immediately adopts the selected policy's action at that state. Revision can therefore interrupt the current action; subsequent completion uses the rate of the newly chosen action. Agents evaluate policies using the stationary long-run payoffs $H(\nu(t))$ at $\bar{\mu}^{\nu(t)}$, as in Section 4, while the state–policy population $\eta(t)$ continues to evolve. The state–policy dynamics with exogenous revision are

$$\dot{\eta}_{s,\pi} = \sum_{s'} X_{s',\pi}(\eta)P_{s's}^{\pi(s')} - X_{s,\pi}(\eta) + \kappa\left(\frac{\gamma_\pi}{m}m_s(\eta) - \eta_{s,\pi}\right), \qquad \gamma \in \Gamma(\nu), \tag{14}$$

where $m_s(\eta) := \sum_\pi \eta_{s,\pi} = \sum_a \mu_{sa}$ is the state mass and the same measurable best-response selection $\gamma(t) \in \Gamma(\nu(t))$ is used for every state–policy pair (s,π). The first two terms describe state transitions for agents using each policy. The last term describes policy revision at the current physical state, which can change the chosen action and hence congestion. Summing over states cancels the physical transition terms and gives

$$\dot{\nu} \in \kappa\big(\Gamma(\nu) - \nu\big). \tag{15}$$

In slow time $\widetilde{t} = \kappa t$, the policy marginal follows the unit-rate best-response inclusion of Section 4, and hence approaches $\mathcal{E}_\Pi^*$ at every positive revision rate. To obtain convergence of the state–policy population, we must also control physical adjustment while these policy masses change. We first set $\kappa = 0$, so the policy distribution ν stays fixed while physical motion and shared congestion remain active. We call the resulting system the *frozen state–policy dynamics*, written here in physical time: $\dot{\eta}_{s,\pi} = \sum_{s'} X_{s',\pi}(\eta)P_{s's}^{\pi(s')} - X_{s,\pi}(\eta)$. Its stationary state–policy completion rates $\bar{X}^\nu$ and population $\bar{\eta}^\nu$ are given by

$$\bar{X}_{s,\pi}^\nu := \bar{\chi}_\pi^\nu q_s^\pi, \qquad \bar{\eta}_{s,\pi}^\nu := \bar{X}_{s,\pi}^\nu\,\partial_{s,\pi(s)}G(\bar{x}^\nu). \tag{16}$$

These rates aggregate to the state–action completion-rate vector, $\bar{x}_{sa}^\nu = \sum_{\pi:\pi(s)=a}\bar{X}_{s,\pi}^\nu$. The policy mass balances give $\sum_s \bar{\eta}_{s,\pi}^\nu = \nu_\pi$, and aggregating $\bar{\eta}^\nu$ through (12) gives exactly the state–action population $\bar{\mu}^\nu$ in (8).

Proposition 10. *Under Assumptions 1 and 4, for every fixed policy marginal $\nu \in \mathcal{M}_m(\Pi)$, system* (14) *with $\kappa = 0$ has a unique global solution from every initial point in $\mathcal{E}_\nu := \{\eta \in \mathcal{M}_m(\mathcal{S}\times\Pi) : \nu(\eta) = \nu\}$. Within $\mathcal{E}_\nu$, the unique equilibrium is $\bar{\eta}^\nu$, and every solution converges to it.*

This extends fixed-policy convergence to populations using several policies and sharing congestion. In particular, it justifies the stationary environment used to define $H(\nu)$. We combine this convergence with the ordinal policy potential to analyze learning when revisions are rare. The target is the nonempty, compact, convex set $\mathcal{E}_{\mathcal{S}\times\Pi}^* := \{\bar{\eta}^\nu : \nu \in \mathcal{E}_\Pi^*\}$: the state–policy population is stationary under the frozen state–policy dynamics, and the policy distribution is an equilibrium of the population game.

Theorem 11 (Convergence as policy revisions become rare)**.** *Under Assumptions 1 and 4, for every $\kappa > 0$,* (14) *has a global absolutely continuous solution from every initial point in $\mathcal{M}_m(\mathcal{S}\times\Pi)$. Furthermore, (i) every such solution η^κ satisfies $\mathrm{dist}(\nu^\kappa(t), \mathcal{E}_\Pi^*) \to 0$, $\Delta(\nu^\kappa(t)) \le e^{-\kappa t}\Delta(\nu^\kappa(0))$, and $M_{\mathrm{BR}}(\nu^\kappa(t)) \le e^{-\kappa t}M_{\mathrm{BR}}(\nu^\kappa(0))$; and (ii) if $\kappa_n \downarrow 0$, $\eta_n^\infty \in \omega(\eta^{\kappa_n})$ are arbitrary omega-limit points of arbitrary solutions, and $\eta_{n_k}^\infty \to \eta^*$, then $\eta^* = \bar{\eta}^{\nu^*}$ for some $\nu^* \in \mathcal{E}_\Pi^*$. Equivalently, $\lim_{\kappa\downarrow 0}\sup_{\eta^\kappa(\cdot)}\sup_{\eta^\infty\in\omega(\eta^\kappa)}\mathrm{dist}(\eta^\infty, \mathcal{E}_{\mathcal{S}\times\Pi}^*) = 0$. Every such cluster point has aggregate state–action population $\bar{\mu}^{\nu^*}$ that forms a stationary MFG equilibrium with an optimal stationary policy by Theorem 9. Along any sequence of omega-limit points η_n^∞ as in (ii), $\lambda(\mu(\eta_n^\infty)) \to (1/\tau_i^*)_{i\in\mathcal{I}}$ and $r\cdot x(\mu(\eta_n^\infty))/m \to g^*$.*

For every fixed revision rate $\kappa > 0$, the policy distribution approaches the population-game equilibrium set $\mathcal{E}_\Pi^*$. As κ tends to zero, the state–policy populations at omega-limit points also approach the stationary populations $\bar{\eta}^\nu$ associated with these policy equilibria. This yields stationary MFG equilibria and their common reward and completion rates. We next consider a model in which agents can revise their policies only when their current actions are completed, so congestion also affects the frequency of revision opportunities.

6 EVENT-DRIVEN STATE–POLICY DYNAMICS

We consider event-driven state–policy dynamics in which an agent completes its current action before reconsidering its policy. This is natural for trips, charging sessions, and service tasks. Congestion now determines both the speed of physical motion and the frequency of revision opportunities. Here $\varepsilon \in (0, 1]$ is the probability of revision at an action completion, whereas $\kappa > 0$ in Section 5 is a Poisson revision rate. The limits $\varepsilon \downarrow 0$ and $\kappa \downarrow 0$, respectively, make revision rare relative to physical transitions.

Using the state–policy completion rates $X_{s,\pi}(\eta)$ from (13), define the policy completion rates and total population completion rate by $\chi_\pi(\eta) := \sum_s X_{s,\pi}(\eta)$ and $\Lambda(\eta) := \sum_\pi \chi_\pi(\eta)$. At stationarity, $\chi(\bar{\eta}^\nu) = \bar{\chi}^\nu$. Upon arrival, the agent keeps its policy with probability $1 - \varepsilon$; with probability ε, it samples a policy from the probability distribution γ/m, $\gamma \in \Gamma(\nu)$, using the same long-run policy payoffs as before. For a common measurable best-response selection $\gamma(t) \in \Gamma(\nu(t))$, the event-driven state–policy dynamics are

$$\dot{\eta}_{s,\pi} = (1-\varepsilon) \sum_{s'} X_{s',\pi}(\eta) P^{\pi(s')}_{s's} - X_{s,\pi}(\eta) + \varepsilon \frac{\gamma_\pi}{m} \sum_{s',a} x_{s'a}(\mu) P^a_{s's}. \tag{17}$$

The last term distributes arrivals from all state–action pairs according to the policy probabilities γ/m. These dynamics conserve mass and preserve the state–policy simplex $\mathcal{M}_m(\mathcal{S} \times \Pi)$. Summing over s gives

$$\dot{\nu}_\pi = \varepsilon \left(\frac{\gamma_\pi}{m} \Lambda(\eta) - \chi_\pi(\eta) \right). \tag{18}$$

The policy marginal now depends on the evolving state–policy population through $\chi_\pi(\eta)$. Policies with higher completion rates receive more revision opportunities. This additional feedback is the main obstacle to transferring the policy-space convergence result to the event-driven model. To separate the total completion rate from its distribution across policies, define event time by $\theta(t) := \int_0^t \Lambda(\eta(t'))\, dt'$, which integrates the population's total completion rate. Positivity and continuity of λ on the compact set $\mathcal{M}_m(\mathcal{I})$ give uniform bounds $0 < \underline{\lambda} \le \lambda_i(\mu) \le \overline{\lambda} < \infty$ for every $\mu \in \mathcal{M}_m(\mathcal{I})$ and $i \in \mathcal{I}$. Consequently $m\underline{\lambda} \le \Lambda(\eta) \le m\overline{\lambda}$, so this change of clock is bi-Lipschitz and preserves limit sets and monotonicity along trajectories. We use the same trajectory symbols in either clock when the time variable is clear, and hats to distinguish event-time trajectories when both clocks appear together. Let $e_\pi(\eta) := \chi_\pi(\eta)/\Lambda(\eta)$ be policy π's share of completion events, and write $e(\eta) := (e_\pi(\eta))_{\pi \in \Pi}$. Dividing (18) by $\Lambda(\eta)$ gives

$$\frac{d\nu_\pi}{d\theta} = \varepsilon \left(\frac{\gamma_\pi}{m} - e_\pi(\eta) \right). \tag{19}$$

In slow time $\widetilde{\theta} = \varepsilon\theta$, this becomes $d\nu_\pi/d\widetilde{\theta} = \gamma_\pi/m - e_\pi(\eta)$. The same ordinal potential nevertheless controls policy revision. Its derivative has the following exact form, even while the state–policy population is adjusting.

Proposition 12. *Under Assumptions 1 and 4, along every absolutely continuous event-driven solution, the following identity holds for almost every event time:*

$$\frac{d}{d\theta}\Phi(\nu) = \varepsilon \left[\max_{\sigma \in \Pi} \log H_\sigma(\nu) - \sum_{\pi \in \Pi} e_\pi(\eta) \log H_\pi(\nu) \right] \ge 0. \tag{20}$$

The bracketed term is zero if and only if $\nu \in \mathcal{E}^*_\Pi$. Hence, for every fixed $\varepsilon \in (0, 1]$, $\operatorname{dist}(\nu(\theta), \mathcal{E}^*_\Pi) \to 0$. In physical time, $\Delta(\nu(t)) \le e^{-\varepsilon\underline{\lambda}t}\Delta(\nu(0))$ for all $t \ge 0$.

Identity (20) ensures policy convergence to $\mathcal{E}^*_\Pi$, but this alone does not guarantee that the state–policy population approaches $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ at a fixed revision probability. At $\bar{\eta}^{\nu^*}$ with $\nu^* \in \mathcal{E}^*_\Pi$, arrival and completion rates agree for every state–policy pair. Revision can keep the policy distribution fixed by balancing inflows and outflows for each policy after summing over states. For the state–policy population to remain stationary, revision must also balance these flows for each state–policy pair. We can construct a two-state periodic trajectory under a time-varying measurable best-response selection at ties (see Remark 37 in Appendix E.4). The next theorem shows that the omega-limit sets of the state–policy trajectories nevertheless approach $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ as revisions become rare, recovering the stationary MFG equilibrium target.

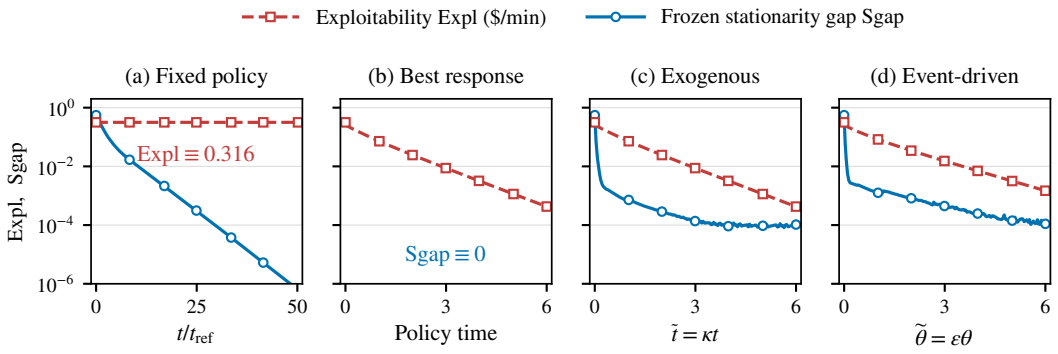


Figure 1: Exploitability and frozen stationarity gap in the Manhattan ride-hailing model. All panels start from the same policy distribution and use common logarithmic vertical scales.

Theorem 13 (Event-driven convergence as policy revisions become rare)**.** *Under Assumptions 1 and 4, for every* $\varepsilon \in (0,1]$, (17) *has a global absolutely continuous solution from every initial point in* $\mathcal{M}_m(\mathcal{S}\times\Pi)$. *Let* $\hat\eta^\varepsilon$ *be any such solution expressed in event time. Every point of* $\omega(\hat\eta^\varepsilon)$ *has policy marginal in* $\mathcal{E}^*_\Pi$. *Moreover,* $\lim_{\varepsilon\downarrow 0}\sup_{\hat\eta^\varepsilon(\cdot)}\sup_{\hat\eta^\infty\in\omega(\hat\eta^\varepsilon)}\operatorname{dist}(\hat\eta^\infty,\mathcal{E}^*_{\mathcal{S}\times\Pi})=0$. *Equivalently, every cluster point of omega-limit points along* $\varepsilon_n\downarrow 0$ *has the form* $\bar\eta^{\nu^*}$ *with* $\nu^*\in\mathcal{E}^*_\Pi$. *Its aggregate state–action population* $\bar\mu^{\nu^*}$ *forms a stationary MFG equilibrium with the optimal stationary policy* $\bar\pi^{\nu^*}$ *given in Theorem 9(i). The limits for individual completion rates and average reward rate in Theorem 11 hold for every sequence* $\hat\eta^\infty_n\in\omega(\hat\eta^{\varepsilon_n})$ *with* $\varepsilon_n\downarrow 0$.

We can further show that, for a singleton policy equilibrium set $\mathcal{E}^*_\Pi=\{\nu^*\}$ consisting of a strict pure equilibrium, every event-driven trajectory converges to the stationary state–policy population $\bar\eta^{\nu^*}$ for every fixed $\varepsilon\in(0,1]$ (Theorem 38 in Appendix E).

Numerical illustration. We illustrate the four dynamics with a ride-hailing model on 66 Manhattan TLC zones, calibrated from March 2026 weekday-morning trips (New York City Taxi and Limousine Commission, 2026). From these data, we calibrate the transition matrices and rewards and construct the convex potential G using observed travel times and a specified congestion delay. Figure 1 compares state–action dynamics under a fixed policy (a), best-response policy dynamics (b), and state–policy dynamics with exogenous (c) or event-driven revision (d). We report policy exploitability Expl (dollars/minute) and the dimensionless frozen stationarity gap Sgap, which measures flow imbalance. With a fixed policy, only the stationarity gap decreases; under policy dynamics, exploitability decreases while the stationarity gap stays zero. Both state–policy systems reduce exploitability and bring the stationarity gap to about 10^{-4} by the end of the plotted window. We use revision rate $\kappa=0.01/t_{\mathrm{ref}}$ in (c) and revision probability $\varepsilon=0.01$ per completion in (d), where $t_{\mathrm{ref}}=24.63$ minutes is the mean cycle time at the reference equilibrium. Appendix G gives the model, clocks, and numerical methods, and shows how less frequent revision reduces the stationarity gap even when exploitability is already small.

Concluding Remarks. We study an MFG with a congestion structure supporting equilibrium computation and convergence of state–action dynamics and best-response learning. We also propose and investigate two state–policy systems with exogenous and event-driven revision. In both systems, policy convergence holds at every fixed admissible positive revision parameter, and the state–action populations associated with omega-limit points approach stationary MFG equilibrium populations as revisions become rare. An important extension is to allow revision opportunities to become rarer over time. We prove convergence of the state–policy population when the revision rate or probability tends to zero (Appendix F). Another extension is to consider other revision protocols; we establish convergence results for specified replicator and Brown–von Neumann–Nash protocols (see Appendices D.5 and E.6). A more interesting direction for future research is model-free learning from observed transitions and rewards. The key difficulty is to control payoff estimation errors while policy updates and population evolution change the environment generating the data. Such an extension would be of particular interest for multi-agent reinforcement learning.

AI USE STATEMENT

Generative AI tools were used to assist with checking mathematical correctness, improving the rigor of proofs, calibrating the numerical model using NYC data, coding the numerical experiments, and revising the manuscript text. The authors reviewed the AI-assisted work, independently checked the mathematical arguments, references, and reported numerical results, and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Complete proofs of all main theoretical claims are provided in Appendices A–E. Appendix G gives the Manhattan calibration, data processing, solver checks, and step-size refinement, and compares the two mechanisms as revisions become rarer. Implementation code and numerical data are provided with the supplementary material.

REFERENCES

Jonatha Anselmi, Bernardo D'Auria, and Neil Walton. Closed queueing networks under congestion: Nonbottleneck independence and bottleneck convergence. *Mathematics of Operations Research*, 38(3):469–491, 2013. doi: 10.1287/moor.1120.0583.

Jean-Pierre Aubin and Halina Frankowska. Heavy viable trajectories of controlled systems. *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire*, 2(5):371–395, 1985. doi: 10.1016/S0294-1449(16)30397-3. URL https://www.numdam.org/item/AIHPC_1985__2_5_371_0/.

Angelos Aveklouris, Maria Vlasiou, and Bert Zwart. A stochastic resource-sharing network for electric vehicle charging. *IEEE Transactions on Control of Network Systems*, 6(3):1050–1061, 2019. doi: 10.1109/TCNS.2019.2915651.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 9780521833783. URL https://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2011. ISBN 9780387709130. doi: 10.1007/978-0-387-70914-7.

Amarjit Budhiraja, Paul Dupuis, Markus Fischer, and Kavita Ramanan. Local stability of Kolmogorov forward equations for finite state nonlinear Markov processes. *Electronic Journal of Probability*, 20(81):1–30, 2015a. doi: 10.1214/EJP.v20-4004.

Amarjit Budhiraja, Paul Dupuis, Markus Fischer, and Kavita Ramanan. Limits of relative entropies associated with weakly interacting particle systems. *Electronic Journal of Probability*, 20(80):1–22, 2015b. doi: 10.1214/EJP.v20-4003.

Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: The fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017. doi: 10.1051/cocv/2016004.

Costas Courcoubetis and Antonis Dimakis. Selfish behavior and resource competition in multi-agent systems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, pp. 529–537. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2025.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/030e65da2b1c944090548d36b244b28d-Abstract.html.

Saeed Hadikhanloo and Francisco J. Silva. Finite mean field games: Fictitious play and convergence to a first order continuous mean field game. *Journal de Mathématiques Pures et Appliquées*, 132:369–397, 2019. doi: 10.1016/j.matpur.2019.02.006.

Josef Hofbauer and William H. Sandholm. Stable games and their dynamics. *Journal of Economic Theory*, 144(4):1665–1693.e4, 2009. doi: 10.1016/j.jet.2009.01.007.

Kuang Huang, Xu Chen, Xuan Di, and Qiang Du. Dynamic driving and routing games for autonomous vehicles on networks: A mean field game approach. *Transportation Research Part C: Emerging Technologies*, 128:103189, 2021. doi: 10.1016/j.trc.2021.103189.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006. doi: 10.4310/CIS.2006.v6.n3.a5.

Noboru Isobe, Kenshi Abe, and Kaito Ariu. Last iterate convergence in monotone mean field games. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20563–20596, 2025. doi: 10.52202/085713-0611.

Frank P. Kelly. On a class of approximations for closed queueing networks. *Queueing Systems*, 4(1): 69–76, 1989. doi: 10.1007/BF01150857.

Nicholas D. Kullman, Martin Cousineau, Justin C. Goodson, and Jorge E. Mendoza. Dynamic ride-hailing with electric vehicles. *Transportation Science*, 56(3):775–794, 2022. doi: 10.1287/trsc.2021.1042.

Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, 45(10): 1572–1589, 2011. doi: 10.1016/j.trb.2011.07.011.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1):229–260, 2007. doi: 10.1007/s11537-007-0657-8.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravishankar Iyer, and Tamer Başar. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36243–36258, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/eb3c8135137c8a60425a0320869ad87e-Abstract-Conference.html.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70 (2):583–601, 2002. doi: 10.1111/1468-0262.00296. URL https://web.stanford.edu/~isegal/envelope.pdf.

Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14:124–143, 1996. doi: 10.1006/game.1996.0044.

Berenice Anne Neumann. Stationary equilibria of mean field games with finite state and action space. *Dynamic Games and Applications*, 10(4):845–871, 2020. doi: 10.1007/s13235-019-00345-9.

Berenice Anne Neumann. Nonlinear Markov chains with finite state space: Invariant distributions and long-term behaviour. *Journal of Applied Probability*, 60(1):30–44, 2023. doi: 10.1017/jpr.2022.23.

Berenice Anne Neumann. A myopic adjustment process for mean field games with finite state and action space. *International Journal of Game Theory*, 53(1):159–195, 2024. doi: 10.1007/s00182-023-00866-z.

New York City Taxi and Limousine Commission. Data Dictionary: High Volume FHV Trip Records. https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_hvfhs.pdf, 2025. March 18, 2025.

New York City Taxi and Limousine Commission. TLC Trip Record Data: High Volume For-Hire Vehicle Records. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2026. January–March 2026 records and taxi-zone geometries; accessed September 7, 2026.

Leonardo Pedroso, Andrea Agazzi, W. P. M. H. Heemels, and Mauro Salazar. Evolutionary dynamics in continuous-time finite-state mean field games – Part I: Equilibria, 2025a. URL https://arxiv.org/abs/2511.01452.

Leonardo Pedroso, Andrea Agazzi, W. P. M. H. Heemels, and Mauro Salazar. Evolutionary dynamics in continuous-time finite-state mean field games – Part II: Stability, 2025b. URL https://arxiv.org/abs/2511.03297.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13199–13213, 2020.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1994. ISBN 9780471619772. doi: 10.1002/9780470316887.

Tim Roughgarden and Éva Tardos. How bad is selfish routing? *Journal of the ACM*, 49(2):236–259, 2002. doi: 10.1145/506147.506153.

David G. Schaeffer and John W. Cain. *Ordinary Differential Equations: Basics and Beyond*, volume 65 of *Texts in Applied Mathematics*. Springer, New York, NY, 2016. doi: 10.1007/978-1-4939-6389-8.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-8328-0. URL https://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11436–11447, 2021.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5571–5580, 2018. URL https://proceedings.mlr.press/v80/yang18d.html.

Enpeng Yuan, Wenbo Chen, and Pascal Van Hentenryck. Reinforcement learning from optimization proxy for ride-hailing vehicle relocation. *Journal of Artificial Intelligence Research*, 75:985–1002, 2022. doi: 10.1613/jair.1.13794.

Sihan Zeng, Sujay Bhatt, Alec Koppel, and Sumitra Ganesh. Learning in herding mean field games: Single-loop algorithm with finite-time convergence analysis. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 343–351, 2025. URL https://proceedings.mlr.press/v258/zeng25a.html.

Chenyu Zhang, Xu Chen, and Xuan Di. Stochastic semi-gradient descent for learning mean field games with population-aware function approximation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/67f30132d98e758f7b4e28c36091d86e-Abstract-Conference.html.

TABLE OF CONTENTS

A Proofs and Technical Details for Section 2 **14**
A.1 Proof of Proposition 2: completion rates . . . 14
A.2 The individual optimization problem as a linear program . . . 15
A.3 Proof of Theorem 5: stationary MFG equilibria . . . 17

B Proofs and Technical Details for Section 3 **19**
B.1 Compactness of trajectories and passage to the limit . . . 19
B.2 Recurrent classes, transient decay, and positivity . . . 20
B.3 Proof of Proposition 6: the fixed-policy Lyapunov function . . . 23
B.4 Proof of Theorem 7: fixed-policy convergence . . . 25

C Proofs and Technical Details for Section 4 **26**
C.1 Optimal unichain policies . . . 26
C.2 Proof of Theorem 8: policy potential and convergence . . . 26
C.3 Other payoff-improving revision protocols . . . 29
C.4 Decomposing stationary completion rates into unichain policies . . . 32
C.5 Proof of Theorem 9: the policy–MFG correspondence . . . 33

D Proofs and Technical Details for Section 5 **34**
D.1 Existence and the policy marginal equation . . . 35
D.2 Proof of Proposition 10: convergence of the frozen state–policy dynamics . . . 35
D.3 Geometry of the stationary target set . . . 40
D.4 Proof of Theorem 11: rare exogenous revisions . . . 40
D.5 State–policy dynamics with exogenous BNN and replicator revision . . . 41

E Proofs and Technical Details for Section 6 **43**
E.1 Existence and marginal equations . . . 43
E.2 Frozen state–policy dynamics in event time . . . 45
E.3 Proof of Proposition 12: event-driven policy convergence . . . 47
E.4 Proof of Theorem 13: rare event-driven revisions . . . 47
E.5 State–policy convergence under a strict pure policy equilibrium . . . 49
E.6 Event-driven state–policy dynamics with BNN and replicator revision . . . 49

F State–policy dynamics with vanishing revision rate and probability **52**

G Numerical Methods and Supplementary Results **55**
G.1 Data and service area . . . 55
G.2 Ride-hailing model . . . 56
G.3 Exploitability and stationarity gap . . . 57
G.4 Computation and experimental setup . . . 57
G.5 Effect of revision frequency . . . 58

A Proofs and Technical Details for Section 2

We first prove Proposition 2. For an individual agent with fixed sojourn times, we formulate the average-reward continuous-time MDP of Section 2 as a linear program and use its dual to prove Theorem 5.

A.1 Proof of Proposition 2: completion rates

Proof of Proposition 2. For $x \geq 0$, convexity and positivity give

$$G(x) - G(0) \geq \nabla G(0) \cdot x \geq \Big(\min_i \partial_i G(0)\Big) \|x\|_1.$$

Hence $\mu = 0$ gives the unique maximizer $x = 0$. For $\mu \neq 0$, the maximization in (2) reduces to the relative interior of the support face

$$x_i > 0 \quad (\mu_i > 0), \qquad x_i = 0 \quad (\mu_i = 0).$$

This face restriction follows because, for any i with $\mu_i > 0$, setting $x_i = 0$ gives objective value $-\infty$. For any i with $\mu_i = 0$, decreasing $x_i > 0$ strictly decreases G without changing the logarithmic term. On this face, the objective has the upper bound

$$\sum_i \mu_i \log x_i - G(x) \leq \|\mu\|_1 \log(1 + \|x\|_1) - \Big(\min_i \partial_i G(0)\Big) \|x\|_1 - G(0).$$

The right-hand side tends to $-\infty$ as $\|x\|_1 \to \infty$. On bounded sets, for any i with $\mu_i > 0$, letting $x_i \downarrow 0$ gives

$$\sum_j \mu_j \log x_j - G(x) \longrightarrow -\infty.$$

Thus every nonempty superlevel set at a finite level is compact. Strict concavity in the variables x_i with $\mu_i > 0$ gives a unique maximizer. The first-order conditions for $\mu_i > 0$, together with $x_i = 0$ when $\mu_i = 0$, give

$$\mu = x \odot \nabla G(x).$$

Conversely, any nonnegative solution of this identity has the same support as μ, since $\nabla G(x) > 0$. On that support, the first-order conditions are sufficient by concavity.

The preceding characterization makes the map $x \mapsto x \odot \nabla G(x)$ a bijection from $\mathbb{R}_+^{\mathcal{I}}$ onto itself, with inverse $x(\mu)$. Its Jacobian on U is

$$D_x[x \odot \nabla G(x)] = \operatorname{diag}(\nabla G(x)) + \operatorname{diag}(x) \nabla^2 G(x). \tag{21}$$

For $x \geq 0$, suppose $D_x[x \odot \nabla G(x)]v = 0$. If $x_i = 0$, the i-th row gives $\partial_i G(x) v_i = 0$, hence $v_i = 0$. Furthermore, summing over the indices i with $x_i > 0$ gives

$$0 = \sum_{i: x_i > 0} \frac{v_i}{x_i} \big(D_x[x \odot \nabla G(x)]v\big)_i = \sum_{i: x_i > 0} \frac{\partial_i G(x)}{x_i} v_i^2 + v^\top \nabla^2 G(x) v \geq \sum_{i: x_i > 0} \frac{\partial_i G(x)}{x_i} v_i^2.$$

Thus $v = 0$, so the Jacobian is nonsingular for every $x \in \mathbb{R}_+^{\mathcal{I}}$, including boundary points where some $x_i = 0$. Fix $\mu^0 \geq 0$. The inverse function theorem at $x(\mu^0)$ provides a C^1 local inverse $\widetilde{x}$ on an open neighborhood of μ^0. Shrink this neighborhood so that $\nabla G(\widetilde{x}) > 0$. For $\mu \geq 0$ in this neighborhood,

$$\widetilde{x}_i(\mu) = \frac{\mu_i}{\partial_i G(\widetilde{x}(\mu))} \geq 0.$$

Uniqueness therefore gives $\widetilde{x}(\mu) = x(\mu)$. The maps $\mu \mapsto 1/\partial_i G(\widetilde{x}(\mu))$ give the corresponding C^1 extensions of λ_i. In particular, $x_i(\mu) = \mu_i \lambda_i(\mu)$, and both maps are locally Lipschitz across support faces. Global Lipschitz continuity follows from compactness of $\mathcal{M}_m(\mathcal{I})$ (Schaeffer & Cain, 2016, Proposition 3.3.2). □

A.2 THE INDIVIDUAL OPTIMIZATION PROBLEM AS A LINEAR PROGRAM

With fixed sojourn times τ, the individual optimization problem is a finite continuous-time MDP. An agent choosing action a in state s holds it for an exponentially distributed time with mean τ_{sa}. Upon completion, it receives reward r_{sa}, moves to state s' with probability $P^a_{ss'}$, and chooses its next action. For a stationary policy π, write Q^π_τ for its generator:

$$Q^\pi_{\tau,(s,a),(s',a')} = \frac{P^a_{ss'}\pi(s',a') - \delta_{(s,a),(s',a')}}{\tau_{sa}}.$$

The LP below follows the average-reward formulation in Puterman (1994, Section 11.4.4, p. 558), with the normalization scaled to population mass m. We give a full proof under Assumption 4 because this formulation is essential for the subsequent results.

Lemma 14. *Under Assumption 4, fix $\tau \in (0,\infty)^{\mathcal{I}}$ and completion rates $\lambda_{sa} = 1/\tau_{sa}$. In the continuous-time MDP of Section 2 with these rates, the optimal average reward is g^*_τ from every initial state, where g^*_τ is $1/m$ times the optimal value of the primal LP:*

$$\begin{aligned} \max_{x\geq 0} \quad & r \cdot x \\ \text{subject to} \quad & \sum_a x_{sa} = \sum_{s',a'} x_{s'a'} P^{a'}_{s's}, \qquad \forall\, s \in \mathcal{S}, \\ & \sum_{s,a} \tau_{sa} x_{sa} = m. \end{aligned} \tag{22}$$

The dual is given by

$$\begin{aligned} \min_{g\in\mathbb{R},\, h\in\mathbb{R}^{\mathcal{S}}} \quad & mg \\ \text{subject to} \quad & r_{sa} - g\tau_{sa} + \sum_{s'} P^a_{ss'} h_{s'} - h_s \leq 0, \qquad \forall\,(s,a) \in \mathcal{I}. \end{aligned} \tag{23}$$

Both programs attain the common optimal value mg^*_τ.

Proof. We first establish LP duality. Here g is the scalar gain variable, measured as reward per agent per unit time, and h_s is the multiplier for the balance equation at state s. The vector h is a state potential (relative value): the constraint is $r_{sa} - g\tau_{sa} \leq h_s - \sum_{s'} P^a_{ss'} h_{s'}$. Adding a common constant to all h_s leaves these constraints unchanged.

Choose a stationary policy and an invariant probability distribution of its embedded state chain, and let ρ be the corresponding invariant state–action probability distribution. Its state marginal is $q_s = \sum_a \rho_{sa}$. It satisfies nonnegativity, $\rho \geq 0$, and normalization, $\sum_{s,a} \rho_{sa} = 1$. Its balance equations are

$$\sum_a \rho_{sa} = \sum_{s',a'} \rho_{s'a'} P^{a'}_{s's}, \qquad s \in \mathcal{S}.$$

Thus $m\rho/(\tau\cdot\rho)$ is feasible for (22). The feasible set is closed. For every feasible x and every $(s,a) \in \mathcal{I}$,

$$0 \leq x_{sa} \leq \frac{m}{\tau_{sa}}.$$

Thus the feasible set is also bounded, hence compact. For a primal-feasible x and a dual-feasible (g,h), normalization and state balance give the duality gap

$$mg - r\cdot x = \sum_{s,a} x_{sa}\left(g\tau_{sa} - r_{sa} + h_s - \sum_{s'} P^a_{ss'} h_{s'}\right) \geq 0.$$

LP duality gives an optimal solution x^*_τ of (22) and an optimal solution (g^*_τ, h^*_τ) of (23), with $r\cdot x^*_\tau = mg^*_\tau$. Fix these solutions. Complementary slackness gives

$$x^*_{\tau,sa}\left(g^*_\tau\tau_{sa} - r_{sa} + h^*_{\tau,s} - \sum_{s'} P^a_{ss'} h^*_{\tau,s'}\right) = 0.$$

We next bound the average reward of every admissible policy. Fix an admissible policy, starting from s_0, and use $N(T)$, S_k, and A_k from Section 2. Count completions of each pair by

$$N_{sa}(T) := \sum_{k=1}^{N(T)} \mathbf{1}_{\{(S_{k-1},A_k)=(s,a)\}}.$$

The completion intensity $1/\tau_{sa}$ gives

$$\mathbb{E}N_{sa}(T) = \frac{1}{\tau_{sa}}\mathbb{E}\int_0^T \mathbf{1}_{\{S_{N(t)}=s,\ A_{N(t)+1}=a\}}\,dt.$$

Multiplying by τ_{sa} and summing over state–action pairs therefore yields

$$\sum_{s,a}\tau_{sa}\mathbb{E}N_{sa}(T) = \mathbb{E}\int_0^T 1\,dt = T.$$

All sums are integrable because $\mathbb{E}N(T) \le T\max_{s,a}\tau_{sa}^{-1}$. The transition probabilities and telescoping give

$$\sum_{s,a}\mathbb{E}N_{sa}(T)\left(\sum_{s'}P^a_{ss'}h^*_{\tau,s'} - h^*_{\tau,s}\right) = \mathbb{E}\sum_{k=1}^{N(T)}(h^*_{\tau,S_k} - h^*_{\tau,S_{k-1}}) = \mathbb{E}h^*_{\tau,S_{N(T)}} - h^*_{\tau,s_0}.$$

Multiplying the dual inequalities by the expected completion counts gives

$$\mathbb{E}\sum_{k=1}^{N(T)} r_{S_{k-1}A_k} = \sum_{s,a} r_{sa}\mathbb{E}N_{sa}(T) \le g^*_\tau T + h^*_{\tau,s_0} - \mathbb{E}h^*_{\tau,S_{N(T)}}.$$

Since the state space is finite,

$$\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{k=1}^{N(T)} r_{S_{k-1}A_k} \le g^*_\tau.$$

This bound includes time-dependent and history-dependent policies.

Finally, we construct a stationary policy from the LP optimizer x^*_τ that achieves this bound from every initial state. Put $\mathcal{S}^*_{\tau,+} = \{s : \sum_a x^*_{\tau,sa} > 0\}$. Choose $\pi^*_\tau(s,a) = x^*_{\tau,sa}/\sum_{a'} x^*_{\tau,sa'}$ on $\mathcal{S}^*_{\tau,+}$, and $\pi^*_\tau(s,a) = 1/|\mathcal{A}|$ elsewhere. The LP balance constraint and the definition of π^*_τ give

$$\begin{aligned}\sum_{s',a'}\tau_{s'a'}x^*_{\tau,s'a'}Q^{\pi^*_\tau}_{\tau,(s',a'),(s,a)} &= \pi^*_\tau(s,a)\sum_{s',a'}x^*_{\tau,s'a'}P^{a'}_{s's} - x^*_{\tau,sa}\\ &= \pi^*_\tau(s,a)\sum_{a'}x^*_{\tau,sa'} - x^*_{\tau,sa} = 0 \qquad \forall\,(s,a)\in\mathcal{I}.\end{aligned}$$

For $s \notin \mathcal{S}^*_{\tau,+}$, both terms in the last line are zero. Since $\tau\cdot x^*_\tau = m$, the vector $\tau\odot x^*_\tau$ is an invariant population of mass m for $Q^{\pi^*_\tau}_\tau$. The completion rates $\sum_a x^*_{\tau,sa}$ at each state form an invariant vector for the embedded state chain, since

$$\sum_{s'}\left(\sum_a x^*_{\tau,s'a}\right)P^{\pi^*_\tau}_{s's} = \sum_{s',a}x^*_{\tau,s'a}P^a_{s's} = \sum_a x^*_{\tau,sa}, \qquad \forall s\in\mathcal{S}. \tag{24}$$

The total completion rate $\sum_{s,a}x^*_{\tau,sa}$ is positive because $\tau\cdot x^*_\tau = m > 0$. Dividing the invariant vector $\left(\sum_a x^*_{\tau,sa}\right)_{s\in\mathcal{S}}$ in (24) by $\sum_{s,a}x^*_{\tau,sa}$ gives an invariant probability distribution with support $\mathcal{S}^*_{\tau,+}$. In a finite chain, the support of an invariant probability distribution is a union of recurrent classes. Since every state in $\mathcal{S}_0$ is transient under π^*_τ, it follows that $\varnothing \ne \mathcal{S}^*_{\tau,+} \subseteq \mathcal{S}_1$.

We now use weak communication to show that every recurrent state class lies in $\mathcal{S}^*_{\tau,+}$. Suppose that $\mathcal{C} \subseteq \mathcal{S}\setminus\mathcal{S}^*_{\tau,+}$ is a recurrent state class of the embedded chain under π^*_τ. For every $s\in\mathcal{C}$, closure of this class and the uniform choice of actions outside $\mathcal{S}^*_{\tau,+}$ give

$$0 = \sum_{s'\notin\mathcal{C}}P^{\pi^*_\tau}_{ss'} = \frac{1}{|\mathcal{A}|}\sum_a\sum_{s'\notin\mathcal{C}}P^a_{ss'}.$$

All summands are nonnegative, so $P^a_{ss'} = 0$ for every $s \in \mathcal{C}$, $s' \notin \mathcal{C}$, and $a \in \mathcal{A}$. Thus every policy keeps the state in $\mathcal{C}$ once it starts there. Its states are recurrent under π^*_τ, so $\mathcal{C} \subseteq \mathcal{S}_1$. Choose $s \in \mathcal{C}$ and $s' \in \mathcal{S}^*_{\tau,+}$. Both belong to $\mathcal{S}_1$, so Assumption 4 gives a deterministic stationary policy under which s' is accessible from s. This is impossible because $s' \notin \mathcal{C}$ and no policy can leave $\mathcal{C}$. Therefore every state outside $\mathcal{S}^*_{\tau,+}$ is transient under π^*_τ.

For each recurrent state class $\mathcal{C}$ of π^*_τ, let $\rho^{\mathcal{C}}$ be the invariant state–action probability distribution supported on $\mathcal{C} \times \mathcal{A}$; it satisfies state balance. If $\rho^{\mathcal{C}}_{sa} > 0$, then $s \in \mathcal{C} \subseteq \mathcal{S}^*_{\tau,+}$ and $\pi^*_\tau(s,a) > 0$, so the definition of π^*_τ gives $x^*_{\tau,sa} > 0$. Complementary slackness for the optimal pair $(x^*_\tau, (g^*_\tau, h^*_\tau))$ therefore makes the dual constraint tight at every such pair:

$$r_{sa} - g^*_\tau \tau_{sa} + \sum_{s'} P^a_{ss'} h^*_{\tau,s'} - h^*_{\tau,s} = 0 \qquad \text{whenever } \rho^{\mathcal{C}}_{sa} > 0.$$

Multiplying by $\rho^{\mathcal{C}}_{sa}$, summing over all pairs, and using state balance to cancel the terms involving h^*_τ, we obtain

$$0 = \sum_{s,a} \rho^{\mathcal{C}}_{sa} \left(r_{sa} - g^*_\tau \tau_{sa} + \sum_{s'} P^a_{ss'} h^*_{\tau,s'} - h^*_{\tau,s} \right) = r \cdot \rho^{\mathcal{C}} - g^*_\tau \, \tau \cdot \rho^{\mathcal{C}}.$$

The stationary average-reward formula (Puterman, 1994, Propositions 11.4.1 and 11.4.7) gives

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi^*_\tau}_s \left[\sum_{k=1}^{N(T)} r_{S_{k-1}A_k} \right] = \frac{\sum_{s',a} \rho^{\mathcal{C}}_{s'a} r_{s'a}}{\sum_{s',a} \rho^{\mathcal{C}}_{s'a} \tau_{s'a}}, \qquad \forall s \in \mathcal{C}. \tag{25}$$

The numerator is the expected reward per completion and the denominator is the expected time per completion, both under $\rho^{\mathcal{C}}$. The denominator is positive, so the preceding identity $r \cdot \rho^{\mathcal{C}} = g^*_\tau \, \tau \cdot \rho^{\mathcal{C}}$ shows that this ratio equals g^*_τ. Every recurrent class has the same gain, and transient states contribute zero to the long-run average. Consequently

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi^*_\tau}_s \left[\sum_{k=1}^{N(T)} r_{S_{k-1}A_k} \right] = g^*_\tau \qquad \forall s \in \mathcal{S}.$$

Together with the upper bound, this shows that the optimal average reward is g^*_τ from every initial state. Both LPs attain the common value $m g^*_\tau$.

□

The objective $r \cdot x$ is the total population reward rate; dividing its optimal value by m gives the reward per agent g^*_τ. The subscript in g^*_τ records the prescribed vector τ; g^* without a subscript denotes the common MFG equilibrium gain in Theorem 5.

From the proof of Lemma 14, we immediately obtain the following.

Corollary 15 (Policy correspondence)**.** *Under the assumptions of Lemma 14, let x^*_τ be any optimizer of* (22)*. Put $\mathcal{S}^*_{\tau,+} = \{s : \sum_a x^*_{\tau,sa} > 0\}$. Define the policy by*

$$\pi^*_\tau(s,a) = \begin{cases} x^*_{\tau,sa} / \sum_{a'} x^*_{\tau,sa'}, & s \in \mathcal{S}^*_{\tau,+}, \\ 1/|\mathcal{A}|, & s \notin \mathcal{S}^*_{\tau,+}. \end{cases}$$

Then π^*_τ is optimal for the prescribed τ from every initial state, and $\tau \odot x^*_\tau$ is an invariant population of mass m for $Q^{\pi^*_\tau}_\tau$.

A.3 Proof of Theorem 5: stationary MFG equilibria

The equilibrium–optimizer correspondence follows the linear-programming and KKT argument of Courcoubetis & Dimakis (2025, Theorem 1). Their program uses linear execution costs and resource constraints with waiting-time multipliers; here the sojourn times are $\nabla G(x)$, and the associated population is $x \odot \nabla G(x)$. For completeness, we give a full proof for our formulation.

Proof of Theorem 5. An invariant state probability distribution q of a deterministic policy π gives a nonzero feasible vector $\widehat{x}_{s,\pi(s)} = q_s$, zero elsewhere. Set $r_{\max} = \max_i r_i$. Since $\partial_i G(0) > 0$ for every i, every nonzero feasible x satisfies

$$m \log(r \cdot x) - G(x) \le m \log(r_{\max} \|x\|_1) - \Big(\min_i \partial_i G(0)\Big) \|x\|_1 - G(0) \longrightarrow -\infty \text{ as } \|x\|_1 \to 0 \text{ or } \infty.$$

The superlevel set at $\widehat{x}$ is compact; continuity gives attainment, and concavity gives a nonempty compact convex solution set $\mathcal{X}^*$. Since the constraints are affine, the refined Slater condition and the KKT theorem apply (Boyd & Vandenberghe, 2004, Sections 5.2.3 and 5.5.3): a feasible x is optimal if and only if some $\vartheta \in \mathbb{R}^{\mathcal{S}}$ satisfies

$$\frac{m r_{sa}}{r \cdot x} - \partial_{sa} G(x) - \vartheta_s + \sum_{s'} P^a_{ss'} \vartheta_{s'} \le 0, \qquad \text{with equality if } x_{sa} > 0. \tag{26}$$

(i) Let (μ^*, π^*) be an equilibrium, $x^* = x(\mu^*)$, and $\tau^* = \nabla G(x^*)$. Population consistency gives

$$x^*_{sa} = \pi^*(s,a) \sum_{s',a'} x^*_{s'a'} P^{a'}_{s's}.$$

Summing over actions yields state balance:

$$\sum_a x^*_{sa} = \sum_{s',a'} x^*_{s'a'} P^{a'}_{s's}.$$

Little's law gives the mass constraint $\tau^* \cdot x^* = m$. Individual optimality and Lemma 14 give the average reward $g^* := g^*_{\tau^*}$. Averaging the reward rate r_{sa}/τ^*_{sa} under the stationary probability distribution μ^*/m yields

$$g^* = \sum_{s,a} \frac{\mu^*_{sa}}{m} \frac{r_{sa}}{\tau^*_{sa}} = \frac{r \cdot x^*}{m}.$$

Thus x^* attains the optimal LP value mg^*. An optimal dual pair (g^*, h^*) has $g^* = r \cdot x^*/m > 0$ and satisfies

$$r_{sa} - g^* \tau^*_{sa} - h^*_s + \sum_{s'} P^a_{ss'} h^*_{s'} \le 0.$$

Complementary slackness gives

$$x^*_{sa} \left(r_{sa} - g^* \tau^*_{sa} - h^*_s + \sum_{s'} P^a_{ss'} h^*_{s'} \right) = 0.$$

Division by g^*, with $\vartheta^* = h^*/g^*$, gives (26) at (x^*, ϑ^*); hence $x^* \in \mathcal{X}^*$.

Conversely, let $x^* \in \mathcal{X}^*$, choose ϑ^* satisfying (26) at x^*, and put $\tau^* = \nabla G(x^*)$, $g^* = r \cdot x^*/m$. Complementary slackness and state balance yield

$$\begin{aligned} 0 &= \sum_{s,a} x^*_{sa} \left(\frac{m r_{sa}}{r \cdot x^*} - \tau^*_{sa} - \vartheta^*_s + \sum_{s'} P^a_{ss'} \vartheta^*_{s'} \right) \\ &= m - \tau^* \cdot x^* + \sum_s \vartheta^*_s \left(\sum_{s',a'} x^*_{s'a'} P^{a'}_{s's} - \sum_a x^*_{sa} \right) = m - \tau^* \cdot x^*. \end{aligned}$$

Thus x^* is feasible for (22) with $\tau = \tau^*$. Multiplying (26) at (x^*, ϑ^*) by g^* shows that $(g^*, g^* \vartheta^*)$ is dual-feasible. Since $mg^* = r \cdot x^*$, weak duality proves LP-optimality. Set $\mu^* = x^* \odot \tau^* = x^* \odot \nabla G(x^*)$. Proposition 2 gives $x(\mu^*) = x^*$ and $\lambda_{sa}(\mu^*) = 1/\tau^*_{sa}$. Corollary 15 supplies a stationary policy π^*, optimal for these sojourn times, with μ^* invariant under $Q^{\pi^*}_{\tau^*}$. Since $Q^{\pi^*}_{\tau^*} = Q^{\pi^*}(\mu^*)$, $(\mu^*, \pi^*) \in \mathcal{E}^*$. The inverse identities $\mu^* = x^* \odot \nabla G(x^*)$ and $x^* = x(\mu^*)$ prove the claimed bijection between $\mathcal{X}^*$ and $\mathcal{E}^*_{\mathcal{I}}$.

(ii) For $x^*, \widetilde{x}^* \in \mathcal{X}^*$ and $0 < \alpha < 1$, write $x^*_\alpha = (1-\alpha)x^* + \alpha \widetilde{x}^*$. Since $x^*_\alpha \in \mathcal{X}^*$,

$$0 = m \left[\log(r \cdot x^*_\alpha) - (1-\alpha) \log(r \cdot x^*) - \alpha \log(r \cdot \widetilde{x}^*) \right] + (1-\alpha) G(x^*) + \alpha G(\widetilde{x}^*) - G(x^*_\alpha).$$

Both summands are nonnegative. Strict concavity of log gives $r \cdot x^* = r \cdot \widetilde{x}^*$. Since the objective is constant on $\mathcal{X}^*$, it follows that

$$G(x^*_\alpha) = G(x^*) = G(\widetilde{x}^*) \qquad (0 \le \alpha \le 1).$$

Differentiating at $\alpha = 0$ gives $\nabla G(x^*) \cdot (\widetilde{x}^* - x^*) = 0$. The supporting-hyperplane inequality now yields, for every $x \in U$,

$$G(x) \ge G(x^*) + \nabla G(x^*) \cdot (x - x^*) = G(\widetilde{x}^*) + \nabla G(x^*) \cdot (x - \widetilde{x}^*).$$

Hence $\nabla G(x^*) \in \partial G(\widetilde{x}^*) = \{\nabla G(\widetilde{x}^*)\}$. Writing $\tau^* = \nabla G(x^*)$, Part (i) gives the common equilibrium rates $\lambda_{sa}(\mu^*) = 1/\tau^*_{sa}$ and gain $g^* = r \cdot x^*/m$.

(iii) Parts (i)–(ii) identify the population set as

$$\mathcal{E}^*_{\mathcal{I}} = \mathrm{diag}(\tau^*)\mathcal{X}^*.$$

Since every $\tau^*_i > 0$, the displayed linear map is invertible. It preserves nonemptiness, compactness, and convexity. □

Corollary 16 (Uniqueness of the equilibrium population)**.** *Under the assumptions of Theorem 5, suppose additionally that G is strictly convex on the feasible set of* (3)*. Then* (3) *has exactly one optimal solution x^*, and there is exactly one state–action population μ^* for which (μ^*, π^*) is a stationary MFG equilibrium for some stationary policy π^*.*

Proof. Theorem 5 guarantees existence of an optimizer. Strict concavity of $x \mapsto m \log(r \cdot x) - G(x)$ gives $\mathcal{X}^* = \{x^*\}$. Theorem 5(i) then yields

$$\mathcal{E}^*_{\mathcal{I}} = \{x^* \odot \nabla G(x^*)\} = \{\mu^*\}.$$

□

B Proofs and Technical Details for Section 3

Fix a stationary mixed unichain policy π. We prove that the state–action dynamics under π converge to the unique stationary population, as stated in Theorem 7. The proof of global convergence has two components: transient mass decays, and a strict Lyapunov function controls motion in the recurrent class. Together, these arguments establish convergence for all initial populations, including those on the boundary.

B.1 Compactness of trajectories and passage to the limit

Here we prove technical lemmas for ODEs that will be used to analyze the learning dynamics. The following lemma applies Arzelà–Ascoli (Teschl, 2012, Theorem 2.18, p. 55) to trajectories of ODEs and their approximations. It provides the compactness needed for the convergence proofs below.

Lemma 17. *Let $I = \mathbb{R}$ or $I = [0, \infty)$, set $I_T = I \cap [-T, T]$ for $T > 0$, and let $0 < T_n \to \infty$. Suppose $K \subset \mathbb{R}^d$ is compact and $u_n : I_{T_n} \to K$ are absolutely continuous with $\|\dot{u}_n(t)\| \le c_u$ almost everywhere, for one constant $0 \le c_u < \infty$.*

(i) There is a subsequence $(u_{n_k})_{k \ge 1}$ converging to a c_u-Lipschitz map $u : I \to K$, with

$$\sup_{t \in I_T} \|u_{n_k}(t) - u(t)\| \longrightarrow 0 \quad (k \to \infty) \qquad \text{for every } T < \infty. \tag{27}$$

(ii) If $f : K \to \mathbb{R}^d$ is continuous and the original sequence (u_n) satisfies, for every $T < \infty$,

$$\sup_{s,t \in I_T} \left\| u_n(t) - u_n(s) - \int_s^t f(u_n(t'))\, dt' \right\| \longrightarrow 0 \quad (n \to \infty), \tag{28}$$

then the limit u in (i) satisfies $u(t) - u(s) = \int_s^t f(u(t'))\, dt'$ for all $s, t \in I$.

Proof. (i) For fixed T and all sufficiently large n, compactness gives $\sup_{t\in I_T}\|u_n(t)\| \le \max_{z\in K}\|z\|$. Absolute continuity gives the common modulus of continuity

$$\|u_n(t)-u_n(s)\| \le \int_{\min(s,t)}^{\max(s,t)} \|\dot{u}_n(t')\|\,dt' \le c_u|t-s|. \tag{29}$$

Successive applications of Arzelà–Ascoli give infinite index sets $\mathcal{N}_1 \supset \mathcal{N}_2 \supset \cdots$, with $T_n \ge j$ for $n \in \mathcal{N}_j$, and limits $u^{(j)} : I_j \to K$ satisfying

$$\lim_{\substack{n\to\infty \\ n\in\mathcal{N}_j}} \sup_{t\in I_j} \|u_n(t)-u^{(j)}(t)\| = 0, \qquad j = 1,2,\ldots.$$

Nestedness gives $u^{(j+1)}|_{I_j} = u^{(j)}$, so define $u|_{I_j} = u^{(j)}$. Choose $n_k \in \mathcal{N}_k$ with $n_{k+1} > n_k$. For any $T < \infty$, choose an integer $j \ge T$. Since $n_k \in \mathcal{N}_j$ for $k \ge j$,

$$\sup_{t\in I_T}\|u_{n_k}(t)-u(t)\| \le \sup_{t\in I_j}\|u_{n_k}(t)-u^{(j)}(t)\| \longrightarrow 0 \qquad (k\to\infty).$$

Closedness of K gives $u(I) \subset K$, and passage to the limit along n_k in (29) gives $\|u(t)-u(s)\| \le c_u|t-s|$.

(ii) Uniform continuity of f gives $\sup_{I_T}\|f(u_{n_k})-f(u)\| \to 0$ as $k\to\infty$. Condition (28) also holds along n_k. Consequently, for $s,t \in I_T$,

$$\begin{aligned}\left\|u(t)-u(s)-\int_s^t f(u(t'))\,dt'\right\| &\le 2\sup_{I_T}\|u-u_{n_k}\| + 2T\sup_{I_T}\|f(u_{n_k})-f(u)\| \\ &\quad + \sup_{a,b\in I_T}\left\|u_{n_k}(b)-u_{n_k}(a)-\int_a^b f(u_{n_k}(t'))\,dt'\right\| \\ &\longrightarrow 0 \quad (k\to\infty).\end{aligned}$$

□

The convergence proofs below use a Lyapunov function L that is nonincreasing along ODE trajectories and decreases strictly away from the equilibrium. The next lemma uses this property and compactness to show that the complete trajectories obtained as limits must be constant at the equilibrium. An ODE solution trajectory is called *complete* if it is defined for all real times.

Lemma 18. *Let F be a locally Lipschitz vector field on an open subset $\mathcal{O}$ of a finite-dimensional affine space. Suppose $L \in C^1(\mathcal{O})$ satisfies $L \ge 0$ and $DL(z)[F(z)] \le 0$, with equality in the latter only at $z^* \in \mathcal{O}$, where $L(z^*) = 0$. Every complete solution $\zeta : \mathbb{R} \to \mathcal{O}$ of $\dot{\zeta} = F(\zeta)$ whose orbit closure is a compact subset of $\mathcal{O}$ is identically equal to z^*.*

Proof. Compactness and monotonicity give the finite limit $L_- := \lim_{t\to-\infty} L(\zeta(t))$. Choose $t_n \to -\infty$ such that $\zeta(t_n) \to z^- \in \mathcal{O}$. Let $\phi_{t'}(z)$ denote the state at time t' of the ODE solution starting from z, so $\phi_0(z) = z$ and $\partial_{t'}\phi_{t'}(z) = F(\phi_{t'}(z))$. Continuous dependence on the initial state gives, for all sufficiently small $t' \ge 0$,

$$L(\phi_{t'}(z^-)) = \lim_n L(\zeta(t_n+t')) = L_- = L(z^-).$$

Differentiation at $t' = 0$ yields $DL(z^-)[F(z^-)] = 0$. By hypothesis, $z^- = z^*$, so $L_- = L(z^*) = 0$. It follows that

$$0 \le L(\zeta(t)) \le L_- = 0 \qquad (t\in\mathbb{R}).$$

Thus $L\circ\zeta$ is constant and $DL(\zeta(t))[F(\zeta(t))] = 0$ for every t. The equality condition again gives $\zeta(t) = z^*$. □

B.2 Recurrent classes, transient decay, and positivity

Here we prove technical lemmas which establish exponential decay of transient mass and uniform positivity in the recurrent class. We first verify the recurrent and transient classes of the state–action

chain stated in Section 3, using the standard finite-chain classification (Puterman, 1994, Secs. A.2–A.3, pp. 588–590).

Recall $p^\pi_{(s,a),(s',a')} = P^a_{ss'}\pi(s',a')$ and $P^\pi_{ss'} = \sum_a \pi(s,a)P^a_{ss'}$. For $(s,a) \in \mathcal{I}^\pi_{\mathrm{rec}}$ and $s' \notin \mathcal{S}^\pi_{\mathrm{rec}}$, closure of $\mathcal{S}^\pi_{\mathrm{rec}}$ gives $0 = P^\pi_{ss'} \geq \pi(s,a)P^a_{ss'}$. Since $\pi(s,a) > 0$, we have $P^a_{ss'} = 0$, so $\mathcal{I}^\pi_{\mathrm{rec}}$ is closed. Starting from (s,a), the first state transition uses action a and has probabilities $P^a_{ss''}$. The next n state transitions follow P^π, and the final action a' is chosen with probability $\pi(s',a')$. Thus, for each integer $n \geq 0$,

$$\left[(p^\pi)^{n+1}\right]_{(s,a),(s',a')} = \pi(s',a') \sum_{s''\in\mathcal{S}} P^a_{ss''}\left[(P^\pi)^n\right]_{s''s'}.$$

For any $i = (s,a)$ and $j = (s',a')$ in $\mathcal{I}^\pi_{\mathrm{rec}}$, choose $s'' \in \mathcal{S}^\pi_{\mathrm{rec}}$ with $P^a_{ss''} > 0$. Irreducibility of P^π on $\mathcal{S}^\pi_{\mathrm{rec}}$ gives an $n \geq 0$ such that

$$\left[(p^\pi)^{n+1}\right]_{ij} \geq \pi(s',a')P^a_{ss''}\left[(P^\pi)^n\right]_{s''s'} > 0.$$

Thus $\mathcal{I}^\pi_{\mathrm{rec}}$ is irreducible. For any $i = (s,a) \in \mathcal{I}^\pi$, the same identity gives

$$\sum_{j\in\mathcal{I}^\pi_{\mathrm{tr}}} \left[(p^\pi)^{n+1}\right]_{ij} = \sum_{s''\in\mathcal{S}} P^a_{ss''} \sum_{s'\in\mathcal{S}^\pi_{\mathrm{tr}}} \left[(P^\pi)^n\right]_{s''s'} \longrightarrow 0 \qquad (n\to\infty),$$

because $\mathcal{S}^\pi_{\mathrm{tr}}$ consists of transient states of P^π. Hence all pairs in $\mathcal{I}^\pi_{\mathrm{tr}}$ are transient, and $\mathcal{I}^\pi_{\mathrm{rec}}$ is the unique recurrent class of the state–action chain. The next lemma quantifies the decay of transient mass.

Lemma 19. *Under Assumption 1, fix a stationary mixed unichain policy π. Let p^π_{tr} be the submatrix of p^π on $\mathcal{I}^\pi_{\mathrm{tr}}$, and put*

$$\widetilde{w}^\pi := (\mathrm{I} - p^\pi_{\mathrm{tr}})^{-1}\mathbf{1}, \qquad \widetilde{W}_\pi(\mu) := \sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}} \widetilde{w}^\pi_i \mu_i,$$

with $\widetilde{W}_\pi \equiv 0$ if $\mathcal{I}^\pi_{\mathrm{tr}} = \varnothing$. Then $\widetilde{w}^\pi_i \geq 1$ for every $i \in \mathcal{I}^\pi_{\mathrm{tr}}$, and along (4)

$$\dot{\widetilde{W}}_\pi = -\sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}} x_i(\mu) \leq -c_{\mathrm{tr},\pi}\widetilde{W}_\pi \tag{30}$$

for some $c_{\mathrm{tr},\pi} > 0$. In particular, every complete solution $\zeta : \mathbb{R} \to \mathcal{M}^\pi_m$ of (4) *satisfies $\widetilde{W}_\pi(\zeta(t)) = 0$ for all $t \in \mathbb{R}$.*

Proof. For $\mathcal{I}^\pi_{\mathrm{tr}} \neq \varnothing$, the transient-block formula (Puterman, 1994, Proposition A.3, p. 593) gives

$$\widetilde{w}^\pi = (\mathrm{I} - p^\pi_{\mathrm{tr}})^{-1}\mathbf{1} = \sum_{n\geq 0}(p^\pi_{\mathrm{tr}})^n\mathbf{1} \geq \mathbf{1}.$$

In particular, $(\mathrm{I} - p^\pi_{\mathrm{tr}})\widetilde{w}^\pi = \mathbf{1}$. Closure of $\mathcal{I}^\pi_{\mathrm{rec}}$ gives $\dot{\mu}_{\mathcal{I}^\pi_{\mathrm{tr}}} = ((p^\pi_{\mathrm{tr}})^\top - \mathrm{I})x_{\mathcal{I}^\pi_{\mathrm{tr}}}$. Using $x_i = \mu_i\lambda_i$ and the uniform rate bounds from Section 6,

$$\dot{\widetilde{W}}_\pi = (\widetilde{w}^\pi)^\top((p^\pi_{\mathrm{tr}})^\top - \mathrm{I})x_{\mathcal{I}^\pi_{\mathrm{tr}}} = -\sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}} x_i \leq -\underline{\lambda}\sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}}\mu_i \leq -\frac{\underline{\lambda}}{\|\widetilde{w}^\pi\|_\infty}\widetilde{W}_\pi.$$

Thus (30) holds with $c_{\mathrm{tr},\pi} = \underline{\lambda}/\|\widetilde{w}^\pi\|_\infty$, and

$$\sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}}\mu_i(t) \leq \widetilde{W}_\pi(\mu(t)) \leq e^{-c_{\mathrm{tr},\pi}t}\widetilde{W}_\pi(\mu(0)) \leq \|\widetilde{w}^\pi\|_\infty e^{-c_{\mathrm{tr},\pi}t}\sum_{i\in\mathcal{I}^\pi_{\mathrm{tr}}}\mu_i(0).$$

For $\mathcal{I}^\pi_{\mathrm{tr}} = \varnothing$, both sides of (30) are zero. For a complete solution ζ, the same estimate starting at $t - n$ and compactness of $\mathcal{M}^\pi_m$ give, for every $t \in \mathbb{R}$,

$$0 \leq \widetilde{W}_\pi(\zeta(t)) \leq e^{-c_{\mathrm{tr},\pi}n}\max_{\mu\in\mathcal{M}^\pi_m}\widetilde{W}_\pi(\mu) \longrightarrow 0 \qquad (n\to\infty). \qquad \square$$

The next lemma proves that, when the transition matrix is irreducible and the rates are bounded above and away from zero, every component of the population vector $\mu(t)$ is bounded below by a fixed positive fraction of the total mass after a fixed time.

Lemma 20. *Let E be a finite set and let $P = (P_{ij})_{i,j\in E}$ be an irreducible stochastic matrix. Suppose $\mu(t) \in \mathbb{R}_+^E$ is absolutely continuous on a time interval and satisfies*

$$\dot{\mu}_j(t) = \sum_{i\in E} \mu_i(t)\lambda_i(t)P_{ij} - \mu_j(t)\lambda_j(t), \qquad j \in E, \tag{31}$$

with conserved mass $\sum_{i\in E} \mu_i(t) = \hat{m} > 0$ and measurable rates obeying

$$0 < \underline{\lambda} \le \lambda_i(t) \le \overline{\lambda} < \infty.$$

Then there exist $T_E < \infty$ and $c_E > 0$, depending only on $P, \underline{\lambda}, \overline{\lambda}$, such that

$$\mu_j(t + T_E) \ge c_E \hat{m} \tag{32}$$

for every $j \in E$ and every t for which the trajectory is defined on $[t, t + T_E]$. In particular, if the trajectory is complete, every component of $\mu(t)$ is bounded below by $c_E\hat{m}$ at every time.

Proof. For fixed t and almost every $t' \ge t$, Equation (31) gives, entrywise,

$$\begin{aligned}\frac{d}{dt'}\left(e^{\overline{\lambda}(t'-t)}\mu_j(t')\right) &= e^{\overline{\lambda}(t'-t)}\left[\sum_{i\in E}\lambda_i(t')P_{ij}\mu_i(t') + (\overline{\lambda} - \lambda_j(t'))\mu_j(t')\right] \\ &\ge \underline{\lambda}e^{\overline{\lambda}(t'-t)}\sum_{i\in E} P_{ij}\mu_i(t').\end{aligned}$$

Integrating this differential inequality from t to $t + T$ gives

$$e^{\overline{\lambda}T}\mu(t + T) \ge \mu(t) + \underline{\lambda}\int_t^{t+T} e^{\overline{\lambda}(t'-t)}P^\top\mu(t')\,dt'.$$

The integral is nonnegative, so $e^{\overline{\lambda}(t'-t)}\mu(t') \ge \mu(t)$ for every $t' \in [t, t + T]$. We now prove by induction that, for every $N \ge 0$,

$$e^{\overline{\lambda}(t'-t)}\mu(t') \ge \sum_{k=0}^{N} \frac{\underline{\lambda}^k(t'-t)^k}{k!}(P^\top)^k\mu(t) \qquad (t' \in [t, t + T]).$$

The case $N = 0$ follows from the preceding bound. Suppose the inequality holds for some $N \ge 0$. Substituting this lower bound into the integral and using the nonnegative entries of $P^\top$, we obtain

$$e^{\overline{\lambda}T}\mu(t + T) \ge \mu(t) + \sum_{k=0}^{N} \frac{\underline{\lambda}^{k+1}}{k!}(P^\top)^{k+1}\mu(t)\int_t^{t+T}(t'-t)^k\,dt' = \sum_{k=0}^{N+1}\frac{(\underline{\lambda}T)^k}{k!}(P^\top)^k\mu(t).$$

Since $T \ge 0$ is arbitrary within the trajectory's domain, this proves the induction step. Letting $N \to \infty$,

$$\mu(t + T) \ge e^{-\overline{\lambda}T}e^{\underline{\lambda}TP^\top}\mu(t).$$

Irreducibility gives, for every $i, j \in E$, some $k \ge 0$ with $(P^k)_{ij} > 0$; hence

$$(e^{\underline{\lambda}P})_{ij} \ge \frac{\underline{\lambda}^k}{k!}(P^k)_{ij} > 0.$$

Take $T_E = 1$ and choose

$$c_E = e^{-\overline{\lambda}}\min_{i,j\in E}(e^{\underline{\lambda}P})_{ij} > 0.$$

Then $\mu_j(t + 1) \ge c_E\sum_{i\in E}\mu_i(t) = c_E\hat{m}$. For a trajectory defined on $[t_0, \infty)$, this gives $\mu_j(t) \ge c_E\hat{m}$ for $t \ge t_0+1$. If the trajectory is complete, its domain is $\mathbb{R}$, so $[t-1, t]$ lies in its domain for every $t \in \mathbb{R}$. Applying the estimate with starting time $t - 1$ therefore gives $\mu_j(t) \ge c_E\sum_{i\in E}\mu_i(t-1) = c_E\hat{m}$ at every real time. □

B.3 Proof of Proposition 6: the fixed-policy Lyapunov function

We first justify the stationary population constructed in Section 3.

Lemma 21 (Invariant probability distribution and population normalization)**.** *Under Assumption 1, fix a stationary mixed unichain policy π. Its unique invariant state probability distribution q^π is positive on $\mathcal{S}^\pi_{\mathrm{rec}}$ and zero on $\mathcal{S}^\pi_{\mathrm{tr}}$. The probability distribution $\rho^\pi_{sa} = q^\pi_s \pi(s,a)$, restricted to $\mathcal{I}^\pi$, is the unique invariant probability distribution of p^π, with support $\mathcal{I}^\pi_{\mathrm{rec}}$. There is a unique $\bar\Lambda^\pi > 0$ such that $\bar x^\pi = \bar\Lambda^\pi \rho^\pi$ satisfies $\bar x^\pi \cdot \nabla G(\bar x^\pi) = m$. Then $\bar\mu^\pi = \bar x^\pi \odot \nabla G(\bar x^\pi)$ is stationary for* (4) *and* $x(\bar\mu^\pi) = \bar x^\pi$.

Proof. Let q^π be the unique invariant probability distribution of P^π (Puterman, 1994, Theorem A.2 and Proposition A.3), and put $\rho^\pi_{sa} = q^\pi_s \pi(s,a)$. For $j = (s', a')$,

$$\sum_i \rho^\pi_i p^\pi_{ij} = \pi(s',a') \sum_s q^\pi_s P^\pi_{ss'} = q^\pi_{s'} \pi(s',a') = \rho^\pi_j.$$

Thus ρ^π is the invariant probability distribution of the state–action chain, positive exactly on $\mathcal{I}^\pi_{\mathrm{rec}}$, and every nonzero stationary completion-rate vector is a positive multiple of ρ^π. For $\xi \geq 0$, define $\varphi_\pi(\xi) = \xi(\rho^\pi)^\top \nabla G(\xi\rho^\pi)$. For $\xi > 0$, convexity gives

$$\varphi'_\pi(\xi) = (\rho^\pi)^\top \nabla G(\xi\rho^\pi) + \xi(\rho^\pi)^\top \nabla^2 G(\xi\rho^\pi)\rho^\pi > 0.$$

Also, $\varphi_\pi(0) = 0$, and monotonicity of the gradient along the ray gives

$$\varphi_\pi(\xi) \geq \xi(\rho^\pi)^\top \nabla G(0) \longrightarrow \infty \quad \text{as } \xi \to \infty.$$

There is therefore a unique $\bar\Lambda^\pi > 0$ with $\varphi_\pi(\bar\Lambda^\pi) = m$. The completion-rate vector $\bar x^\pi = \bar\Lambda^\pi \rho^\pi$ therefore satisfies stationarity

$$\sum_{i \in \mathcal{I}^\pi} \bar x^\pi_i p^\pi_{ij} = \bar x^\pi_j \quad (j \in \mathcal{I}^\pi), \tag{33}$$

and the required population normalization

$$\sum_{i \in \mathcal{I}^\pi} \bar x^\pi_i \partial_i G(\bar x^\pi) = m. \tag{34}$$

Consequently $\bar\mu^\pi = \bar x^\pi \odot \nabla G(\bar x^\pi)$ has mass m, $x(\bar\mu^\pi) = \bar x^\pi$, and $\dot\mu|_{\bar\mu^\pi} = (p^\pi)^\top \bar x^\pi - \bar x^\pi = 0$. □

Proof of Proposition 6. (i) Write $x = x(\mu)$. Both x and $\bar x^\pi$ vanish outside $\mathcal{I}^\pi_{\mathrm{rec}}$, and $\bar x^\pi_i > 0$ on $\mathcal{I}^\pi_{\mathrm{rec}}$. Substituting the objective value at its maximizer into (5) and using $\mu_i = x_i \partial_i G(x)$ from (1) gives

$$\begin{aligned} L_\pi(\mu) &= \sum_{i \in \mathcal{I}^\pi_{\mathrm{rec}}} \mu_i \log x_i - G(x) - \sum_{i \in \mathcal{I}^\pi_{\mathrm{rec}}} \mu_i \log \bar x^\pi_i + G(\bar x^\pi) \\ &= G(\bar x^\pi) - G(x) + \sum_{i \in \mathcal{I}^\pi_{\mathrm{rec}}} \partial_i G(x) x_i \log \frac{x_i}{\bar x^\pi_i} \\ &= G(\bar x^\pi) - G(x) - \nabla G(x) \cdot (\bar x^\pi - x) + \sum_{i \in \mathcal{I}^\pi_{\mathrm{rec}}} \partial_i G(x) \left(x_i \log \frac{x_i}{\bar x^\pi_i} - x_i + \bar x^\pi_i \right). \end{aligned}$$

Convexity gives $G(\bar x^\pi) - G(x) - \nabla G(x) \cdot (\bar x^\pi - x) \geq 0$. For each $i \in \mathcal{I}^\pi_{\mathrm{rec}}$, the remaining bracket satisfies

$$x_i \log \frac{x_i}{\bar x^\pi_i} - x_i + \bar x^\pi_i = \bar x^\pi_i \left(\frac{x_i}{\bar x^\pi_i} \log \frac{x_i}{\bar x^\pi_i} - \frac{x_i}{\bar x^\pi_i} + 1 \right) \geq 0,$$

because $z \log z - z + 1 \geq 0$ for $z \geq 0$, with equality exactly at $z = 1$. Since every $\partial_i G(x) > 0$, these calculations show $L_\pi(\mu) \geq 0$ and characterize equality:

$$L_\pi(\mu) = 0 \iff x(\mu) = \bar x^\pi \iff \mu = \bar x^\pi \odot \nabla G(\bar x^\pi) = \bar\mu^\pi.$$

Here the convention $0 \log 0 = 0$ makes the calculation valid when $\mu_i = 0$, hence $x_i = 0$, for some $i \in \mathcal{I}^\pi_{\mathrm{rec}}$; the corresponding bracket then equals $\bar x^\pi_i > 0$.

(ii) Write $x = x(\mu)$. All sums in this part are over $\mathcal{I}^\pi_{\mathrm{rec}}$, where $x_i, \bar x^\pi_i > 0$; for $i \notin \mathcal{I}^\pi_{\mathrm{rec}}$, the population masses μ_i and completion rates $x_i, \bar x^\pi_i$ remain zero. By Proposition 2, the optimizer map $\mu \mapsto x(\mu)$

in (2) is C^1. The product and chain rules, followed by Little's law (1), give the envelope identity (Milgrom & Segal, 2002, Theorem 1) explicitly:

$$\begin{aligned} df^* &= d\left(\sum_i \mu_i \log x_i - G(x)\right) \\ &= \sum_i \left(\log x_i \, d\mu_i + \mu_i \, d(\log x_i)\right) - dG(x) \\ &= \sum_i \left(\log x_i \, d\mu_i + \frac{\mu_i}{x_i} \, dx_i\right) - \sum_i \partial_i G(x) \, dx_i \\ &= \sum_i \log x_i \, d\mu_i + \sum_i \left(\frac{\mu_i}{x_i} - \partial_i G(x)\right) dx_i \\ &= \sum_i \log x_i \, d\mu_i + \sum_i \left(\frac{x_i \partial_i G(x)}{x_i} - \partial_i G(x)\right) dx_i \\ &= \sum_i \log x_i \, d\mu_i. \end{aligned}$$

The differential identity above therefore implies that f^* is C^2 on this face. The recurrent class is closed, so $\sum_j p^\pi_{ij} = 1$ for every $i \in \mathcal{I}^\pi_{\mathrm{rec}}$. Substituting (4), renaming the index in the outflow sum, and using this row-sum identity gives

$$\begin{aligned} \dot{L}_\pi &= \sum_j \dot{\mu}_j \left(\log x_j - \log \bar{x}^\pi_j\right) \\ &= \sum_j \left(\sum_i x_i p^\pi_{ij} - x_j\right) \log \frac{x_j}{\bar{x}^\pi_j} \\ &= \sum_{i,j} x_i p^\pi_{ij} \log \frac{x_j}{\bar{x}^\pi_j} - \sum_i x_i \log \frac{x_i}{\bar{x}^\pi_i} \\ &= \sum_{i,j} x_i p^\pi_{ij} \log \frac{x_j}{\bar{x}^\pi_j} - \sum_i x_i \left(\sum_j p^\pi_{ij}\right) \log \frac{x_i}{\bar{x}^\pi_i} \\ &= \sum_{i,j} x_i p^\pi_{ij} \left(\log \frac{x_j}{\bar{x}^\pi_j} - \log \frac{x_i}{\bar{x}^\pi_i}\right) \\ &= \sum_{i,j} x_i p^\pi_{ij} \log \frac{x_j/\bar{x}^\pi_j}{x_i/\bar{x}^\pi_i}. \end{aligned}$$

For any stochastic matrix p with stationary positive vector $\bar{x}$, we have $\sum_j p_{ij} = 1$ and $\sum_i \bar{x}_i p_{ij} = \bar{x}_j$. Consequently,

$$\begin{aligned} &\sum_{i,j} x_i p_{ij} \left(\frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} - 1\right) \\ &\quad = \sum_{i,j} x_i p_{ij} \frac{x_j \bar{x}_i}{\bar{x}_j x_i} - \sum_{i,j} x_i p_{ij} \\ &\quad = \sum_{i,j} \bar{x}_i p_{ij} \frac{x_j}{\bar{x}_j} - \sum_i x_i \sum_j p_{ij} \\ &\quad = \sum_j \frac{x_j}{\bar{x}_j} \left(\sum_i \bar{x}_i p_{ij}\right) - \sum_i x_i \\ &\quad = \sum_j \frac{x_j}{\bar{x}_j} \bar{x}_j - \sum_i x_i = \sum_j x_j - \sum_i x_i = 0. \end{aligned}$$

Subtracting this zero sum from the logarithmic sum and collecting terms gives

$$\begin{aligned}
&\sum_{i,j} x_i p_{ij} \log \frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} \\
&\quad = \sum_{i,j} x_i p_{ij} \log \frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} - \sum_{i,j} x_i p_{ij} \left(\frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} - 1 \right) \\
&\quad = \sum_{i,j} x_i p_{ij} \left(\log \frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} - \frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} + 1 \right) \\
&\quad = -\sum_{i,j} x_i p_{ij} \left(\frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} - 1 - \log \frac{x_j/\bar{x}_j}{x_i/\bar{x}_i} \right) \le 0.
\end{aligned} \tag{35}$$

The last inequality follows from $z - 1 - \log z \ge 0$ for $z > 0$, with equality exactly at $z = 1$, and the nonnegative weights $x_i p_{ij}$. Thus the parenthesized expression is zero exactly when $x_j/\bar{x}_j = x_i/\bar{x}_i$. Applying this identity with $p = p^\pi$ and $\bar{x} = \bar{x}^\pi$, irreducibility gives

$$\dot{L}_\pi = 0 \iff \frac{x_i}{\bar{x}_i^\pi} = \frac{x_j}{\bar{x}_j^\pi} \text{ whenever } p_{ij}^\pi > 0 \iff x = k\bar{x}^\pi \quad (k > 0).$$

Both $x = k\bar{x}^\pi$ and $\bar{x}^\pi$ satisfy the mass normalization $x \cdot \nabla G(x) = m$. Uniqueness of the scaling in Lemma 21 therefore gives $k = 1$. Thus $x = \bar{x}^\pi$ and $\mu = \bar{\mu}^\pi$; conversely, at $\bar{\mu}^\pi$ the derivative is zero. □

B.4 Proof of Theorem 7: fixed-policy convergence

The following proof uses an argument based on LaSalle's invariance principle (Teschl, 2012, Theorem 6.14, p. 202), adapted to our setting. We need to first show that every limiting complete trajectory has positive mass at every recurrent state–action pair, i.e., $\mu_i > 0$ for every $i \in \mathcal{I}_{\mathrm{rec}}^\pi$, and zero mass at all transient pairs, as required by the Lyapunov derivative formula in Proposition 6(ii).

Proof of Theorem 7. Proposition 2 gives a locally Lipschitz vector field. Row stochasticity gives mass conservation:

$$\sum_j \dot{\mu}_j = \sum_i x_i \Big(\sum_j p_{ij}^\pi - 1 \Big) = 0.$$

If the population mass μ_j at pair j is zero, then $x_j = 0$, so $\dot{\mu}_j = \sum_i x_i p_{ij}^\pi \ge 0$. Thus $\mathcal{M}_m^\pi$ is invariant and every solution is global and unique. The ω-limit set $\omega(\mu)$ is nonempty by compactness of $\mathcal{M}_m^\pi$. Lemma 19 gives $\mu_i(t) \to 0$ for every $i \in \mathcal{I}_{\mathrm{tr}}^\pi$, so

$$\omega(\mu) \subset \{\mu' \in \mathcal{M}_m^\pi : \mu_i' = 0 \ (i \in \mathcal{I}_{\mathrm{tr}}^\pi)\}.$$

Fix $\mu^\infty \in \omega(\mu)$, choose $t_n \ge n$ with $\mu(t_n) \to \mu^\infty$, and set $\mu_n(t) = \mu(t_n + t)$ on $[-n, n]$. The estimate

$$\|\dot{\mu}_n(t)\|_1 \le \sum_j \sum_i x_i(\mu_n) p_{ij}^\pi + \sum_j x_j(\mu_n) \le 2m\bar{\lambda}$$

and Lemma 17 give a complete solution ζ and a subsequence μ_{n_k} that converges uniformly on every bounded time interval: for every $T > 0$,

$$\sup_{|t| \le T} \|\mu_{n_k}(t) - \zeta(t)\| \longrightarrow 0 \qquad (k \to \infty).$$

At $t = 0$, this convergence gives $\zeta(0) = \lim_k \mu_{n_k}(0) = \lim_k \mu(t_{n_k}) = \mu^\infty$. For any fixed $t \in \mathbb{R}$, the shifted trajectories are defined at t for all sufficiently large k, and

$$\mu(t_{n_k} + t) = \mu_{n_k}(t) \longrightarrow \zeta(t), \qquad t_{n_k} + t \longrightarrow \infty.$$

Thus $\zeta(t) \in \omega(\mu)$ by the definition of the omega-limit set. Since $t \in \mathbb{R}$ was arbitrary, $\zeta(\mathbb{R}) \subset \omega(\mu)$; hence $\zeta_i(t) = 0$ for every $i \in \mathcal{I}_{\mathrm{tr}}^\pi$ and $t \in \mathbb{R}$. The closed recurrent state–action class is $\mathcal{I}_{\mathrm{rec}}^\pi = \{(s, a) : s \in \mathcal{S}_{\mathrm{rec}}^\pi,\ \pi(s, a) > 0\}$. The recurrent state–action transition matrix $p_{\mathrm{rec}}^\pi := (p_{ij}^\pi)_{i,j \in \mathcal{I}_{\mathrm{rec}}^\pi}$ is

stochastic and irreducible, and $\sum_{i\in\mathcal{I}^\pi_{\rm rec}}\zeta_i(t)=m$. Apply Lemma 20 with $P=p^\pi_{\rm rec}$ to $(\zeta_i)_{i\in\mathcal{I}^\pi_{\rm rec}}$ on every interval $[t-1,t]$ to obtain

$$\zeta_i(t)\ge c_{\mathcal{I}^\pi_{\rm rec}}m>0\qquad(i\in\mathcal{I}^\pi_{\rm rec},\ t\in\mathbb{R}).$$

This uniform lower bound makes $\overline{\zeta(\mathbb{R})}$ a compact subset of the recurrent-face interior, i.e., the set

$$\{\mu'\in\mathcal{M}^\pi_m:\mu'_i>0\ (i\in\mathcal{I}^\pi_{\rm rec}),\quad \mu'_i=0\ (i\in\mathcal{I}^\pi_{\rm tr})\}.$$

Proposition 6 and Lemma 18 imply $\zeta\equiv\bar\mu^\pi$, so $\mu^\infty=\zeta(0)=\bar\mu^\pi$. Since μ^∞ was arbitrary, $\omega(\mu)=\{\bar\mu^\pi\}$; compactness then gives $\mu(t)\to\bar\mu^\pi$. Continuity gives $x(\mu(t))\to x(\bar\mu^\pi)=\bar x^\pi$. A stationary solution is constant, so the convergence just proved forces it to equal $\bar\mu^\pi$, proving uniqueness. □

C Proofs and Technical Details for Section 4

In this section, we first show that the unichain family in (6) contains an optimal policy for every fixed positive vector of sojourn times τ. We then establish regularity and concavity of Φ in (10), prove quantitative best-response convergence and point convergence for replicator dynamics, and relate policy equilibria (9) to stationary MFG pairs and (3).

C.1 Optimal unichain policies

Lemma 22. *Under Assumption 4, let σ be a deterministic stationary policy and let $\mathcal{C}\subseteq\mathcal{S}$ be one recurrent class of states in its embedded chain. There exists a deterministic stationary policy π that agrees with σ on $\mathcal{C}$ and for which $\mathcal{C}$ is the unique recurrent class. Consequently, the family Π in* (6) *is nonempty, and for every positive vector of sojourn times τ, the individual agent's average-reward continuous-time MDP admits an optimal deterministic unichain policy. Equivalently,*

$$g^*_\tau=\max_{\pi\in\Pi}\frac{r^\pi}{\sum_s q^\pi_s\tau_{s,\pi(s)}},$$

where q^π and r^π are as in Section 4.

Proof. By Puterman (1994, Theorem 8.3.2(a)), the policy σ can be modified outside $\mathcal{C}$ to obtain a deterministic stationary policy π that agrees with σ on $\mathcal{C}$ and has $\mathcal{C}$ as its unique recurrent class. Applying this result to any deterministic stationary policy and one of its recurrent classes gives $\Pi\neq\varnothing$. Choose an optimal primal–dual pair $(x^*_\tau,(g^*_\tau,h^*_\tau))$ from Lemma 14, and put $\mathcal{S}^*_{\tau,+}=\{s:\sum_a x^*_{\tau,sa}>0\}$. For each $s\in\mathcal{S}^*_{\tau,+}$, choose and fix exactly one action $\sigma(s)\in\{a\in\mathcal{A}:x^*_{\tau,sa}>0\}$, and choose any one action at each remaining state. This defines a deterministic stationary policy σ. For $s'\notin\mathcal{S}^*_{\tau,+}$, state balance gives

$$\sum_{s,a}x^*_{\tau,sa}P^a_{ss'}=\sum_{a'}x^*_{\tau,s'a'}=0.$$

Nonnegativity implies $P^a_{ss'}=0$ whenever $x^*_{\tau,sa}>0$. Thus $\mathcal{S}^*_{\tau,+}$ is closed for the embedded state chain under σ. This set is nonempty because $\tau\cdot x^*_\tau=m>0$, so its finiteness implies that it contains a recurrent state class $\mathcal{C}$ of that chain. Theorem 8.3.2(a) gives a deterministic stationary policy π^*_τ that agrees with σ on $\mathcal{C}$ and has $\mathcal{C}$ as its unique recurrent class, by modifying the action choices outside $\mathcal{C}$. For every $s\in\mathcal{C}$, we therefore have $x^*_{\tau,s\pi^*_\tau(s)}=x^*_{\tau,s\sigma(s)}>0$. Applying the complementary-slackness and stationary-reward argument from the proof of Lemma 14 to this recurrent class shows that π^*_τ has average reward g^*_τ from every initial state, since all states outside $\mathcal{C}$ are transient. Lemma 14 bounds every policy's gain by g^*_τ, proving optimality and the maximum formula. □

C.2 Proof of Theorem 8: policy potential and convergence

The next lemma gives the regularity of $\bar\chi^\nu$, defined in (7), needed to differentiate the policy potential Φ defined in (10), including when some policy masses vanish.

Lemma 23. *Under Assumptions 1 and 4, for every $\nu^0 \in \mathbb{R}^{\Pi}_+$, the map $\nu \mapsto \bar{\chi}^\nu$ in (7) is the restriction of a C^1 map on an open neighborhood of ν^0. In particular, there are a neighborhood radius $c_{\mathrm{rad}} > 0$ and a Lipschitz constant $c_{\mathrm{Lip}} > 0$ such that, for all $\nu, \nu' \in \mathbb{R}^{\Pi}_+$ with $\|\nu - \nu^0\| < c_{\mathrm{rad}}$ and $\|\nu' - \nu^0\| < c_{\mathrm{rad}}$,*

$$\|\bar{\chi}^\nu - \bar{\chi}^{\nu'}\| \le c_{\mathrm{Lip}} \|\nu - \nu'\|.$$

This bound also applies when $\nu_\pi = 0 < \nu'_\pi$ for some $\pi \in \Pi$.

Proof. On $\{\chi : \mathbf{R}\chi \in U\}$, the function $\chi \mapsto G(\mathbf{R}\chi)$ is convex because

$$\nabla^2_\chi [G(\mathbf{R}\chi)] = \mathbf{R}^\top \nabla^2 G(\mathbf{R}\chi)\mathbf{R} \succeq 0.$$

Its partial derivatives are positive on $\chi \ge 0$:

$$\frac{\partial}{\partial \chi_\pi} G(\mathbf{R}\chi) = \sum_i \rho_i^\pi \partial_i G(\mathbf{R}\chi) > 0,$$

because $\rho^\pi \ge 0$ and $\sum_i \rho_i^\pi = 1$. The domain is open and convex. Proposition 2 applied to $\chi \mapsto G(\mathbf{R}\chi)$ gives the asserted local C^1 extensions. □

The next lemma uses viability for controlled systems (Aubin & Frankowska, 1985), i.e., the existence of a measurable control $\gamma(t) \in \Gamma(\ell(u(t)))$ and a corresponding solution satisfying the state constraint $u(t) \in K$ for all $t \ge 0$. The lemma applies to the policy dynamics in Section 4 and both state–policy systems in Sections 5 and 6; in each case, a common control $\gamma(t)$ governs all components.

Lemma 24 (Existence for affine revision fields)**.** *Suppose H is continuous on $\mathcal{M}_m(\Pi)$. Let $K \subset \mathbb{R}^d$ be nonempty, compact, and convex, and let $\ell : K \to \mathcal{M}_m(\Pi)$, $F_0 : K \to \mathbb{R}^d$, and $\mathsf{B} : K \to \mathbb{R}^{d \times \Pi}$ be continuous. Suppose some $\Delta t_0 > 0$ satisfies*

$$u + \Delta t\big[F_0(u) + \mathsf{B}(u)\gamma\big] \in K \quad \text{for all } u \in K, \quad \gamma \in \Gamma(\ell(u)), \quad 0 < \Delta t \le \Delta t_0. \tag{36}$$

Then from every $u_0 \in K$ there is a global absolutely continuous trajectory $u : [0, \infty) \to K$ and a measurable selection $\gamma(t) \in \Gamma(\ell(u(t)))$ such that $\dot{u} = F_0(u) + \mathsf{B}(u)\gamma$ almost everywhere.

Proof. Suppose $u_n \to u$, $\gamma_n \to \gamma$, and $\gamma_n \in \Gamma(\ell(u_n))$. Since $\mathcal{M}_m(\Pi)$ is closed, $\gamma \in \mathcal{M}_m(\Pi)$. If $\gamma_\pi > 0$, then $\gamma_{n,\pi} > 0$ for all sufficiently large n, so the definition of Γ in (9) gives

$$H_\pi(\ell(u_n)) \ge H_\sigma(\ell(u_n)) \qquad \text{for every } \sigma \in \Pi.$$

Continuity of ℓ and H gives $H_\pi(\ell(u)) \ge H_\sigma(\ell(u))$ for every $\sigma \in \Pi$. Thus every policy with $\gamma_\pi > 0$ is a best response at $\ell(u)$, so $\gamma \in \Gamma(\ell(u))$. This proves that the control map $u \mapsto \Gamma(\ell(u))$ has a closed graph. For each $u \in K$, the set of admissible controls $\Gamma(\ell(u))$ is a nonempty compact convex subset of $\mathcal{M}_m(\Pi)$. The closed-graph property and compactness of $\mathcal{M}_m(\Pi)$ therefore imply that $u \mapsto \Gamma(\ell(u))$ is upper semicontinuous on K. For $u \in K$, define the contingent tangent cone by

$$\mathrm{Tan}_K(u) = \left\{ v \in \mathbb{R}^d : \liminf_{\delta \downarrow 0} \frac{\mathrm{dist}(u + \delta v, K)}{\delta} = 0 \right\}.$$

The Euler-step condition (36) makes the distance in this definition zero for $v = F_0(u) + \mathsf{B}(u)\gamma$ and every sufficiently small $\delta > 0$. Thus

$$F_0(u) + \mathsf{B}(u)\gamma \in \mathrm{Tan}_K(u), \qquad u \in K, \quad \gamma \in \Gamma(\ell(u)).$$

By the viability theorem for controlled systems (Aubin & Frankowska, 1985, Theorem 2.1), for every $u_0 \in K$ there exist $T > 0$, an absolutely continuous trajectory $u : [0, T] \to K$, and a measurable map $\gamma : [0, T] \to \mathcal{M}_m(\Pi)$ such that $u(0) = u_0$ and

$$\gamma(t) \in \Gamma(\ell(u(t))), \qquad \dot{u}(t) = F_0(u(t)) + \mathsf{B}(u(t))\gamma(t) \quad \text{for almost every } t \in [0, T].$$

The same vector $\gamma(t)$ is used in all components of this equation. Moreover,

$$\|\dot{u}\| \le \max_{u \in K} \|F_0(u)\| + m \max_{u \in K} \|\mathsf{B}(u)\| =: c < \infty.$$

Suppose a maximal continuation is defined on $[0, T_*)$ with $T_* < \infty$. For $0 \le s < t < T_*$, the bound above gives

$$\|u(t) - u(s)\| \le \int_s^t \|\dot{u}(t')\|\, dt' \le c(t-s).$$

Thus $u(t)$ is Cauchy as $t \uparrow T_*$, and closedness of K gives $u_* := \lim_{t\uparrow T_*} u(t) \in K$. Applying local existence at u_* and concatenating the resulting trajectory and measurable control with the original ones at T_* extends the solution beyond T_*, contradicting maximality. Hence the solution can be continued for all $t \ge 0$. □

Proof of Theorem 8. (i) Lemma 23 gives local C^1 extensions of $\bar{\chi} = \bar{\chi}^\nu$. Write $\tau^\pi = \tau^\pi(\bar{\chi})$. The identities $\nu_\pi = \bar{\chi}_\pi \tau^\pi$ and $H_\pi = r^\pi/\tau^\pi$ give

$$\Phi(\nu) = \sum_\pi \nu_\pi \log H_\pi(\nu) - G(\mathbf{R}\bar{\chi}). \tag{37}$$

These identities hold on the local extension, with $\tau^\pi > 0$, so

$$\begin{aligned} d\Phi &= \sum_\pi \log H_\pi \, d\nu_\pi - \sum_\pi \frac{\nu_\pi}{\tau^\pi}\, d\tau^\pi - \sum_\pi \tau^\pi \, d\bar{\chi}_\pi \\ &= \sum_\pi \log H_\pi \, d\nu_\pi - d\Big(\sum_\pi \bar{\chi}_\pi \tau^\pi\Big) = \sum_\pi (\log H_\pi - 1)\, d\nu_\pi. \end{aligned}$$

Thus Φ is the restriction of a C^2 function on an open neighborhood of each $\nu^0 \in \mathbb{R}_+^\Pi$, including points with $\nu_\pi^0 = 0$ for some $\pi \in \Pi$. With $\partial_\pi \Phi := \partial\Phi/\partial\nu_\pi$, the partial derivatives satisfy

$$\operatorname{sgn}(\partial_\pi \Phi - \partial_\sigma \Phi) = \operatorname{sgn}(\log H_\pi - \log H_\sigma) = \operatorname{sgn}(H_\pi - H_\sigma).$$

(ii) For each π, the logarithmic perspective $\nu_\pi \log(\chi_\pi/\nu_\pi)$ is jointly concave in $(\nu_\pi, \chi_\pi) \in \mathbb{R}_+^2$, with the term interpreted as 0 when $\nu_\pi = 0 \le \chi_\pi$ and as $-\infty$ when $\chi_\pi = 0 < \nu_\pi$. Since $\chi \mapsto G(\mathbf{R}\chi)$ is convex,

$$\sum_\pi \left[\nu_\pi \log \frac{\chi_\pi}{\nu_\pi} + \nu_\pi \log r^\pi\right] - G(\mathbf{R}\chi)$$

is jointly concave, and its partial maximum Φ is concave (Boyd & Vandenberghe, 2004, Sections 3.2.5–3.2.6). The first-order characterization on the simplex gives

$$\begin{aligned} \nu \in \arg\max \Phi &\iff \nabla\Phi(\nu)\cdot(\nu' - \nu) \le 0 \quad \forall \nu' \in \mathcal{M}_m(\Pi) \\ &\iff m \max_\pi \log H_\pi(\nu) - \sum_\pi \nu_\pi \log H_\pi(\nu) = 0 \\ &\iff \nu \in \mathcal{E}_\Pi^*. \end{aligned}$$

Since Φ is continuous on the compact simplex $\mathcal{M}_m(\Pi)$, its maximizer set $\mathcal{E}_\Pi^*$ is nonempty and compact. Concavity of Φ and convexity of $\mathcal{M}_m(\Pi)$ imply that $\mathcal{E}_\Pi^*$ is convex. Recall that $\Delta(\nu) = \Phi^* - \Phi(\nu)$ is the potential gap, where $\Phi^* = \max_{\nu' \in \mathcal{M}_m(\Pi)} \Phi(\nu')$. For $\nu \in \mathcal{M}_m(\Pi)$ and $\nu^* \in \mathcal{E}_\Pi^*$, concavity of Φ and the formula $\partial_\pi \Phi = \log H_\pi - 1$ give

$$0 \le \Delta(\nu) \le \nabla\Phi(\nu)\cdot(\nu^* - \nu) \le M_{\mathrm{BR}}(\nu). \tag{38}$$

(iii) Apply Lemma 24 with $K = \mathcal{M}_m(\Pi)$, $\ell(\nu) = \nu$, $F_0(\nu) = -\nu$, and $\mathsf{B}(\nu) = \mathrm{I}$. The payoff map H is continuous by (i), and the Euler-step condition holds because $(1-\Delta t)\nu + \Delta t\gamma \in \mathcal{M}_m(\Pi)$ for $0 < \Delta t \le 1$ and $\gamma \in \Gamma(\nu)$. Thus a global solution exists from every initial point. Fix any such solution and set $\gamma(t) = \nu(t) + \dot{\nu}(t) \in \Gamma(\nu(t))$ almost everywhere. Mass conservation, (i), and (38) give

$$\dot{\Phi} = \sum_\pi (\gamma_\pi - \nu_\pi) \log H_\pi = m \max_\pi \log H_\pi - \sum_\pi \nu_\pi \log H_\pi = M_{\mathrm{BR}} \ge \Delta \ge 0. \tag{39}$$

Thus $\dot{\Delta} = -M_{\mathrm{BR}} \le -\Delta$ and $\Delta(\nu(t)) \le e^{-t}\Delta(\nu(0))$. Compactness of $\mathcal{M}_m(\Pi)$ makes $\omega(\nu)$ nonempty. For any $\nu^\infty \in \omega(\nu)$, choose $t_n \to \infty$ such that $\nu(t_n) \to \nu^\infty$. Continuity and the gap bound imply $\Delta(\nu^\infty) = \lim_n \Delta(\nu(t_n)) = 0$. Since $\mathcal{E}_\Pi^* = \{\nu : \Delta(\nu) = 0\}$ on the simplex, $\omega(\nu) \subseteq \mathcal{E}_\Pi^*$.

We next show that the best-response gap satisfies $\frac{d}{dt} M_{\mathrm{BR}}(\nu(t)) \le -M_{\mathrm{BR}}(\nu(t))$ along solutions of $\dot\nu \in \Gamma(\nu) - \nu$, following the argument in Hofbauer & Sandholm (2009, Theorem 5.3). Each $t \mapsto \log H_\pi(\nu(t))$ is locally absolutely continuous. An absolutely continuous function has derivative zero almost everywhere on its zero set (Brezis, 2011, Exercise 8.10, p. 239). Applying this fact to each difference $t \mapsto \log H_\pi(\nu(t)) - \log H_\sigma(\nu(t))$, $\pi, \sigma \in \Pi$, and using finiteness of Π, gives, outside one common null set,

$$H_\pi(\nu(t)) = H_\sigma(\nu(t)) \quad \Longrightarrow \quad \frac{d}{dt} \log H_\pi(\nu(t)) = \frac{d}{dt} \log H_\sigma(\nu(t)). \tag{40}$$

Write $\log H(\nu) := (\log H_\pi(\nu))_{\pi\in\Pi}$, and let $D_\nu \log H(\nu)$ denote its Jacobian with respect to ν. Since $\gamma \in \Gamma(\nu)$ is supported on maximizing policies and has mass m, (40) and $D_\nu \log H(\nu) = \nabla^2\Phi(\nu)$ give

$$m \frac{d}{dt} \max_\pi \log H_\pi(\nu(t)) = \gamma^\top \nabla^2\Phi(\nu)\dot\nu \quad \text{a.e.}$$

Consequently, using $\dot\nu = \gamma - \nu$ and (39),

$$\dot M_{\mathrm{BR}} = (\gamma - \nu)^\top \nabla^2\Phi(\nu)\dot\nu - \dot\nu \cdot \log H(\nu) = \dot\nu^\top \nabla^2\Phi(\nu)\dot\nu - M_{\mathrm{BR}} \le -M_{\mathrm{BR}}. \tag{41}$$

Here $\nabla^2\Phi \preceq 0$, also at boundary points by continuity. Integration proves $M_{\mathrm{BR}}(\nu(t)) \le e^{-t} M_{\mathrm{BR}}(\nu(0))$.

It remains to verify global asymptotic stability. Since M_{BR} is continuous and vanishes exactly on $\mathcal{E}^*_\Pi$,

$$\min_{\mathrm{dist}(\nu,\mathcal{E}^*_\Pi)\ge\delta} M_{\mathrm{BR}}(\nu) > 0 \quad (\delta > 0 \text{ whenever this set is nonempty}).$$

Decay of M_{BR} and this positive minimum give $\mathrm{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$. For stability, uniform continuity makes M_{BR} uniformly small near $\mathcal{E}^*_\Pi$. Since $M_{\mathrm{BR}}(\nu(t)) \le M_{\mathrm{BR}}(\nu(0))$, trajectories $\nu(t)$ of the best-response inclusion $\dot\nu \in \Gamma(\nu) - \nu$ starting sufficiently close to $\mathcal{E}^*_\Pi$ cannot reach $\{\nu : \mathrm{dist}(\nu, \mathcal{E}^*_\Pi) \ge \delta\}$. □

Remark 25. The set convergence in Theorem 8(iii) cannot generally be replaced by convergence of every best-response solution to one point. Take one state, two actions, $m = 1$, $G(x) = x_1 + x_2$, and $r_1 = r_2 = 1$. Then $H_1 = H_2 = 1$ everywhere and $\mathcal{E}^*_\Pi = \mathcal{P}(\Pi)$. For $0 < c_{\mathrm{amp}} < 1/(2\sqrt{2})$, set

$$\nu_1(t) = \tfrac{1}{2} + c_{\mathrm{amp}} \sin t, \qquad \gamma_1(t) = \tfrac{1}{2} + c_{\mathrm{amp}}(\sin t + \cos t), \qquad \nu_2 = 1 - \nu_1, \quad \gamma_2 = 1 - \gamma_1.$$

Both vectors lie in the simplex, $\gamma(t) \in \Gamma(\nu(t))$, and $\dot\nu = \gamma - \nu$. This is a nonconstant periodic solution consisting entirely of equilibria.

C.3 Other payoff-improving revision protocols

Taking logarithms preserves the best responses and equilibrium set of the population game. By Theorem 8(i), the game with payoffs $\log H$ has the concave exact potential $\Phi(\nu) + \sum_\pi \nu_\pi$, whose gradient is $(\log H_\pi(\nu))_{\pi\in\Pi}$. On $\mathcal{M}_m(\Pi)$, this potential differs from Φ by the constant m. The logarithmic potential also controls replicator and Brown–von Neumann–Nash revision dynamics evaluated at the original payoffs H.

Corollary 26. *Under Assumptions 1 and 4, consider the following dynamics on $\mathcal{M}_m(\Pi)$:*

(i) best response,

$$\dot\nu \in \Gamma(\nu) - \nu;$$

(ii) replicator,

$$\dot\nu_\pi = \nu_\pi(H_\pi - \bar H),$$

where $\bar H(\nu) := \frac{1}{m}\sum_\sigma \nu_\sigma H_\sigma(\nu)$;

(iii) Brown–von Neumann–Nash (BNN),

$$\dot\nu_\pi = [H_\pi - \bar H]^+ - \frac{\nu_\pi}{m} \sum_\sigma [H_\sigma - \bar H]^+,$$

where $[z]^+ := \max\{z, 0\}$, and payoffs are evaluated at ν.

Each of these dynamics admits a global absolutely continuous solution from every initial point of $\mathcal{M}_m(\Pi)$; for replicator and BNN dynamics this solution is unique. Along every such solution $\nu(t)$, $\Phi(\nu(t))$ is nondecreasing. For best-response and BNN dynamics,

$$\operatorname{dist}(\nu(t), \mathcal{E}_\Pi^*) \longrightarrow 0. \tag{42}$$

Equivalently, $\omega(\nu) \subseteq \mathcal{E}_\Pi^*$. For a replicator trajectory $\nu(\cdot) : [0,\infty) \to \mathcal{M}_m(\Pi)$, write $\mathcal{E}_0(\nu)$ for the invariant face determined by its initial support:

$$\mathcal{E}_0(\nu) := \{\nu' \in \mathcal{M}_m(\Pi) : \nu'_\pi = 0 \text{ whenever } \nu_\pi(0) = 0\}.$$

Then $\nu(t) \to \nu^\infty$ for some $\nu^\infty \in \arg\max_{\nu' \in \mathcal{E}_0(\nu)} \Phi(\nu')$. If $\nu_\pi(0) > 0$ for every $\pi \in \Pi$, then $\mathcal{E}_0(\nu) = \mathcal{M}_m(\Pi)$ and $\nu^\infty \in \mathcal{E}_\Pi^*$, although the limit may lie on the boundary. If some policies have zero initial mass, the limit need not belong to $\mathcal{E}_\Pi^*$, that is, (42) *can fail: those policies remain unused even if they offer higher payoffs at the limit.*

Proof. (i) Best-response dynamics. Theorem 8 gives a global absolutely continuous solution from every initial point, nondecrease of Φ along every solution, and convergence to $\mathcal{E}_\Pi^*$ in the sense of (42).

(ii) Replicator dynamics. The replicator field is locally Lipschitz and satisfies $\sum_\pi \dot\nu_\pi = \sum_\pi \nu_\pi H_\pi - m\bar H = 0$. At $\nu_\pi = 0$, it satisfies $\dot\nu_\pi = 0$. Thus the compact simplex is invariant, and ODE existence and continuation give a unique global solution (Teschl, 2012, Theorem 2.2 and Corollary 2.15). The identity $\partial_\pi \Phi = \log H_\pi - 1$ and conservation of mass give $\dot\Phi = \sum_\pi \dot\nu_\pi \log H_\pi$. Using $\bar H = m^{-1}\sum_\sigma \nu_\sigma H_\sigma$ and $\sum_\sigma \nu_\sigma = m$ gives

$$\dot\Phi = \sum_\pi \nu_\pi (H_\pi - \bar H)\log H_\pi = \frac{1}{m}\sum_{\pi,\sigma} \nu_\pi \nu_\sigma (H_\pi - H_\sigma)\log H_\pi.$$

Exchanging the dummy indices π and σ rewrites the last expression as $-m^{-1}\sum_{\pi,\sigma}\nu_\pi\nu_\sigma(H_\pi - H_\sigma)\log H_\sigma$. Averaging these two expressions yields

$$\dot\Phi = \frac{1}{2m}\sum_{\pi,\sigma}\nu_\pi\nu_\sigma(H_\pi - H_\sigma)(\log H_\pi - \log H_\sigma) \geq 0. \tag{43}$$

Equality holds exactly when the payoffs coincide on $\operatorname{supp}\nu$. Thus Φ is nondecreasing and increases strictly away from replicator rest points. Equality of payoffs on the support alone does not rule out higher payoffs at unused policies. We next prove convergence to a single maximizer of Φ on the initial invariant face.

Fix a replicator trajectory ν. For each π, the coefficient $t \mapsto H_\pi(\nu(t)) - \bar H(\nu(t))$ is continuous and bounded, because H is continuous on the compact simplex. Multiplying the replicator equation $\dot\nu_\pi(t) = \nu_\pi(t)[H_\pi(\nu(t)) - \bar H(\nu(t))]$ by its integrating factor gives

$$\frac{d}{dt}\left[\nu_\pi(t)\exp\left(-\int_0^t [H_\pi(\nu(t')) - \bar H(\nu(t'))]\,dt'\right)\right] = 0.$$

Evaluating the constant at $t = 0$ yields, for every finite $t \geq 0$,

$$\nu_\pi(t) = \nu_\pi(0)\exp\left(\int_0^t [H_\pi(\nu(t')) - \bar H(\nu(t'))]\,dt'\right). \tag{44}$$

The exponential factor is finite and strictly positive. Hence $\nu_\pi(t) > 0$ if $\nu_\pi(0) > 0$, and $\nu_\pi(t) = 0$ if $\nu_\pi(0) = 0$. We first consider full-support initialization, so all components remain positive at every finite time.

Courcoubetis & Dimakis (2025, Theorem 2) prove convergence using relative entropy for replicator dynamics with logarithmic fitness, provided the initial support contains the support of an equilibrium. Here the replicator equation uses the original payoffs H, while $\partial_\pi\Phi = \log H_\pi - 1$. We adapt their relative-entropy argument by adding a multiple of the potential gap Δ; the estimates below control the resulting arithmetic–geometric mean discrepancy. Positivity and continuity of the finitely many payoff functions on the compact simplex give a positive minimum $c_{H,\min}$ and a finite maximum

$c_{H,\max}$, taken over all $\nu \in \mathcal{M}_m(\Pi)$ and $\pi \in \Pi$. Thus $0 < c_{H,\min} \le H_\pi(\nu) \le c_{H,\max} < \infty$. For an arbitrary fixed $\nu^* \in \mathcal{E}^*_\Pi$, use the relative entropy

$$\mathrm{KL}(\nu^* \| \nu) = \sum_{\pi : \nu^*_\pi > 0} \nu^*_\pi \log \frac{\nu^*_\pi}{\nu_\pi}.$$

Set $c_{\mathrm{ent}} = c^2_{H,\max}/(2c^2_{H,\min})$ and $L^{\mathrm{rep}}_{\nu^*}(\nu) := \mathrm{KL}(\nu^* \| \nu) + c_{\mathrm{ent}} \Delta(\nu)$. The mean-value theorem applied to (43) gives

$$\dot\Phi \ge \frac{1}{c_{H,\max}} \sum_\pi \nu_\pi (H_\pi - \bar H)^2.$$

To compare the arithmetic and geometric means, write $H_{\mathrm{geom}} = \exp(m^{-1} \sum_\pi \nu_\pi \log H_\pi)$. Taylor expansion of $\log H_\pi$ at $\bar H$, with cancellation of the linear terms, gives

$$0 \le \log \bar H - \log H_{\mathrm{geom}} \le \frac{1}{2m c^2_{H,\min}} \sum_\pi \nu_\pi (H_\pi - \bar H)^2.$$

Using $1 - H_{\mathrm{geom}}/\bar H \le \log(\bar H / H_{\mathrm{geom}})$, and then the preceding lower bound on $\dot\Phi$, we obtain

$$m(\bar H - H_{\mathrm{geom}}) \le m \bar H (\log \bar H - \log H_{\mathrm{geom}}) \le c_{\mathrm{ent}} \dot\Phi.$$

Concavity of Φ, applied at ν as in (38), and Jensen's inequality yield

$$\sum_\pi \frac{\nu^*_\pi}{m} H_\pi \ge \exp\left(\sum_\pi \frac{\nu^*_\pi}{m} \log H_\pi \right) \ge H_{\mathrm{geom}} e^{\Delta/m}.$$

Since $H_{\mathrm{geom}} \ge c_{H,\min}$,

$$\begin{aligned} \frac{d}{dt} \mathrm{KL}(\nu^* \| \nu) &= m \bar H - \sum_\pi \nu^*_\pi H_\pi \\ &\le m(\bar H - H_{\mathrm{geom}}) - m H_{\mathrm{geom}} (e^{\Delta/m} - 1) \\ &\le c_{\mathrm{ent}} \dot\Phi - c_{H,\min} \Delta. \end{aligned}$$

Thus

$$\dot L^{\mathrm{rep}}_{\nu^*} \le -c_{H,\min} \Delta \le 0. \tag{45}$$

Nonnegativity of $L^{\mathrm{rep}}_{\nu^*}$ gives $\int_0^\infty \Delta(\nu(t))\, dt < \infty$. Since Δ is nonincreasing, $\Delta(\nu(t)) \to 0$. Compactness and continuity of Δ give an accumulation point $\nu^\infty \in \mathcal{E}^*_\Pi$, with $\nu(t_n) \to \nu^\infty$. The estimate (45) holds for every equilibrium, so we may now apply it with $\nu^* = \nu^\infty$. The relative entropy uses only components where $\nu^\infty_\pi > 0$, so

$$\mathrm{KL}(\nu^\infty \| \nu(t_n)) = \sum_{\pi : \nu^\infty_\pi > 0} \nu^\infty_\pi \log \frac{\nu^\infty_\pi}{\nu_\pi(t_n)} \longrightarrow 0.$$

Thus $L^{\mathrm{rep}}_{\nu^\infty}(\nu(t_n)) \to 0$. Its nonnegativity and monotonicity imply $L^{\mathrm{rep}}_{\nu^\infty}(\nu(t)) \to 0$ for the whole trajectory. The equal-mass Pinsker inequality now gives

$$\|\nu(t) - \nu^\infty\|_1^2 \le 2m\, \mathrm{KL}(\nu^\infty \| \nu(t)) \le 2m L^{\mathrm{rep}}_{\nu^\infty}(\nu(t)) \to 0.$$

For a replicator trajectory ν with arbitrary initial point, let $\Pi_0 := \{\pi \in \Pi : \nu_\pi(0) > 0\}$. Formula (44) gives $\nu_\pi(t) = 0$ for $\pi \notin \Pi_0$ and $\nu_\pi(t) > 0$ for $\pi \in \Pi_0$ at every finite time. Thus the trajectory remains in $\mathcal{E}_0(\nu)$ and has positive mass at every policy in Π_0. The restriction of Φ to $\mathcal{E}_0(\nu)$ is concave and satisfies $\partial_\pi \Phi = \log H_\pi - 1$ for $\pi \in \Pi_0$. The preceding entropy argument therefore applies with all policy sums restricted to Π_0, with Φ^* replaced by $\Phi^*_0 := \max_{\nu' \in \mathcal{E}_0(\nu)} \Phi(\nu')$, and with Δ replaced by $\Delta_0(\nu') := \Phi^*_0 - \Phi(\nu')$ for $\nu' \in \mathcal{E}_0(\nu)$. It yields convergence to a single maximizer ν^∞ of Φ on $\mathcal{E}_0(\nu)$. This maximizer need not maximize Φ on the full simplex, because policies outside Π_0 are excluded from the comparison.

(iii) Brown–von Neumann–Nash dynamics. The BNN field is locally Lipschitz because H is locally Lipschitz and the positive-part map is Lipschitz. Its components sum to zero on $\mathcal{M}_m(\Pi)$, and at $\nu_\pi = 0$ it satisfies $\dot\nu_\pi = [H_\pi - \bar H]^+ \ge 0$. The compact simplex is therefore invariant, so ODE

existence and continuation again give a unique global solution (Teschl, 2012, Theorem 2.2 and Corollary 2.15). Conservation of mass and $\partial_\pi \Phi = \log H_\pi - 1$ give $\dot{\Phi} = \sum_\pi \dot{\nu}_\pi \log H_\pi$. Jensen's inequality gives

$$\frac{1}{m} \sum_\pi \nu_\pi \log H_\pi \leq \log \bar{H}.$$

Substitution of the BNN field therefore yields

$$\dot{\Phi} = \sum_\pi [H_\pi - \bar{H}]^+ \left(\log H_\pi - \frac{1}{m} \sum_\sigma \nu_\sigma \log H_\sigma \right) \geq \sum_\pi [H_\pi - \bar{H}]^+ \log \frac{H_\pi}{\bar{H}} \geq 0. \tag{46}$$

Every term in the last sum is positive when $H_\pi > \bar{H}$ and zero otherwise. Thus $\dot{\Phi} = 0$ for BNN if and only if $H_\pi \leq \bar{H}$ for every π. Since $\bar{H}$ is the population-weighted average payoff, this condition forces $H_\pi = \bar{H}$ whenever $\nu_\pi > 0$. It is therefore equivalent to $\nu \in \mathcal{E}^*_\Pi$: every used policy maximizes the payoff over all policies.

It remains to deduce convergence to $\mathcal{E}^*_\Pi$. Because the simplex is compact and invariant, each BNN trajectory has a nonempty omega-limit set. Continuity of Φ on the compact simplex makes it bounded above. Since $\Phi(\nu(t))$ is nondecreasing, it has a finite limit $\Phi_\infty := \lim_{t\to\infty} \Phi(\nu(t))$. For any omega-limit point ν^∞, choose $t_n \to \infty$ such that $\nu(t_n) \to \nu^\infty$. Continuity gives

$$\Phi(\nu^\infty) = \lim_{n\to\infty} \Phi(\nu(t_n)) = \Phi_\infty.$$

Thus Φ is constant on the omega-limit set, as also follows from Teschl (2012, Theorem 6.15, p. 202) applied to $-\Phi$. By invariance of that set, the BNN solution starting from any of its points stays in it and has constant Φ. Differentiating at the initial time gives $\dot{\Phi} = 0$ at that point. The preceding characterization of $\dot{\Phi} = 0$ therefore places the omega-limit set in $\mathcal{E}^*_\Pi$. If (42) failed, there would be a sequence $t_n \to \infty$ and a $\delta > 0$ such that $\operatorname{dist}(\nu(t_n), \mathcal{E}^*_\Pi) \geq \delta$. Compactness would yield a convergent subsequence whose limit lies both in the omega-limit set and at distance at least δ from $\mathcal{E}^*_\Pi$, a contradiction. □

C.4 Decomposing stationary completion rates into unichain policies

To recover every MFG equilibrium population from a policy distribution, we need a representation of every vector of stationary completion rates. The following lemma supplies it by identifying the extreme points of the polytope of invariant state–action probability distributions of embedded chains under stationary policies.

Lemma 27. *Under Assumption 4, let $\mathcal{Q}$ be the polytope of invariant state–action probability distributions of embedded chains under stationary policies,*

$$\mathcal{Q} := \left\{ \rho \geq 0 : \sum_{s,a} \rho_{sa} = 1, \quad \sum_a \rho_{sa} = \sum_{s',a} \rho_{s'a} P^a_{s's} \ \forall s \right\}.$$

Every extreme point of $\mathcal{Q}$ is the invariant state–action probability distribution ρ^π of a deterministic unichain policy $\pi \in \Pi$. Consequently

$$\mathcal{Q} = \operatorname{co}\{\rho^\pi : \pi \in \Pi\}, \tag{47}$$

where co *denotes the convex hull.*

Proof. Let ρ be extreme, and put $q_s = \sum_a \rho_{sa}$, $\mathcal{C}_q = \{s : q_s > 0\}$, $\mathcal{I}_\rho = \{(s,a) : \rho_{sa} > 0\}$. For $s' \notin \mathcal{C}_q$, balance gives

$$\sum_{s,a} \rho_{sa} P^a_{ss'} = q_{s'} = 0.$$

Thus $P^a_{ss'} = 0$ for every $(s,a) \in \mathcal{I}_\rho$. Restrict the balance equations to the variables ρ_{sa} with $(s,a) \in \mathcal{I}_\rho$. The rows for states outside $\mathcal{C}_q$ are zero. For every remaining column, the sum of the coefficients in the rows indexed by $\mathcal{C}_q$ is

$$\sum_{s' \in \mathcal{C}_q} \left(\mathbf{1}_{\{s'=s\}} - P^a_{ss'} \right) = 1 - 1 = 0 \qquad ((s,a) \in \mathcal{I}_\rho).$$

Thus these rows have rank at most $|\mathcal{C}_q| - 1$. Adding the normalization row raises the rank to at most $|\mathcal{C}_q|$. If $|\mathcal{I}_\rho| > |\mathcal{C}_q|$, there is a nonzero ρ', supported on $\mathcal{I}_\rho$, that preserves normalization, $\sum_{s,a} \rho'_{sa} = 0$, and satisfies the homogeneous balance equations

$$\sum_a \rho'_{sa} = \sum_{s',a} \rho'_{s'a} P^a_{s's}.$$

For sufficiently small $\delta > 0$, both $\rho + \delta\rho'$ and $\rho - \delta\rho'$ belong to $\mathcal{Q}$. The decomposition

$$\rho = \tfrac{1}{2}(\rho + \delta\rho') + \tfrac{1}{2}(\rho - \delta\rho')$$

contradicts extremality. Since $|\mathcal{I}_\rho| \geq |\mathcal{C}_q|$, equality holds: exactly one action is selected at each state of $\mathcal{C}_q$.

The induced deterministic chain has positive invariant state probabilities q_s on $\mathcal{C}_q$. If $\mathcal{C}_q$ had two or more recurrent classes, restriction to those classes would give a nontrivial convex decomposition

$$\rho = \sum_{\text{recurrent classes } \mathcal{C}'} \left(\sum_{s \in \mathcal{C}'} q_s \right) \frac{\rho \mathbf{1}_{\mathcal{C}' \times \mathcal{A}}}{\sum_{s \in \mathcal{C}'} q_s},$$

again contradicting extremality. Hence $\mathcal{C}_q$ is one recurrent class. Lemma 22 extends the selected actions to $\pi \in \Pi$ with invariant state–action probability distribution $\rho = \rho^\pi$. Writing $\operatorname{ext} \mathcal{Q}$ for the set of extreme points, and using that every $\rho^\pi \in \mathcal{Q}$ and $\mathcal{Q}$ is a compact polytope,

$$\mathcal{Q} = \operatorname{co}(\operatorname{ext} \mathcal{Q}) \subseteq \operatorname{co}\{\rho^\pi : \pi \in \Pi\} \subseteq \mathcal{Q}.$$

□

C.5 Proof of Theorem 9: the policy–MFG correspondence

We first construct a stationary MFG pair from each policy equilibrium $\nu^* \in \mathcal{E}^*_\Pi$, by showing that its aggregate completion-rate vector solves (22). Conversely, Lemma 27 decomposes the completion-rate vector of any stationary MFG pair into unichain policies and thereby gives a policy equilibrium with the same population vector. Parts (ii)–(iii) then identify the common equilibrium values and their limits along policy dynamics.

Proof of Theorem 9. (i) Fix $\nu^* \in \mathcal{E}^*_\Pi$. Since each ρ^π is balanced,

$$\sum_a \bar{x}^{\nu^*}_{sa} = \sum_{s',a'} \bar{x}^{\nu^*}_{s'a'} P^{a'}_{s's}.$$

Little's law for each policy, $\nu^*_\pi = \bar{\chi}^{\nu^*}_\pi \tau^\pi(\bar{\chi}^{\nu^*})$, gives $\nabla G(\bar{x}^{\nu^*}) \cdot \bar{x}^{\nu^*} = \sum_\pi \nu^*_\pi = m$; the identity $\bar{\mu}^{\nu^*} = \bar{x}^{\nu^*} \odot \nabla G(\bar{x}^{\nu^*})$ and uniqueness in Proposition 2 give $x(\bar{\mu}^{\nu^*}) = \bar{x}^{\nu^*}$. Thus $\bar{x}^{\nu^*}$ is feasible for (22) with $\tau^* = \nabla G(\bar{x}^{\nu^*})$. Nash optimality yields

$$r \cdot \bar{x}^{\nu^*} = \sum_\pi \bar{\chi}^{\nu^*}_\pi r^\pi = \sum_\pi \nu^*_\pi H_\pi(\nu^*) = m \max_\pi H_\pi(\nu^*).$$

By Lemma 22, $\max_\pi H_\pi(\nu^*) = g^*_{\tau^*}$ for these fixed sojourn times, so $\bar{x}^{\nu^*}$ solves (22). Here $\tau^* = \nabla G(x(\bar{\mu}^{\nu^*}))$, so Corollary 15 supplies a stationary policy $\bar{\pi}^{\nu^*}$, optimal at the population $\bar{\mu}^{\nu^*}$ from every initial state, whose rows satisfy (11). State balance gives

$$\bar{\pi}^{\nu^*}(s,a) \sum_{s',a'} \bar{x}^{\nu^*}_{s'a'} P^{a'}_{s's} = \bar{x}^{\nu^*}_{sa}.$$

The identity holds also at zero-throughput states, where both sides vanish. It is exactly the consistency condition, so $(\bar{\mu}^{\nu^*}, \bar{\pi}^{\nu^*}) \in \mathcal{E}^*$.

Conversely, take $(\mu^*, \pi^*) \in \mathcal{E}^*$, and set $x^* = x(\mu^*)$, $\Lambda^* = \sum_{s,a} x^*_{sa} > 0$. Lemma 27 supplies coefficients $\alpha_\pi \geq 0$ for $\pi \in \Pi$, with $\sum_{\pi \in \Pi} \alpha_\pi = 1$, such that

$$\rho^* := x^*/\Lambda^* = \sum_{\pi \in \Pi} \alpha_\pi \rho^\pi. \tag{48}$$

By Theorem 5(ii), $\nabla G(x^*) = \tau^*$. With $\tau^{\pi,*} = \rho^\pi \cdot \tau^*$, individual optimality at the sojourn times τ^* gives $r^\pi - g^*\tau^{\pi,*} \le 0$ for every π. The decomposition satisfies

$$\sum_\pi \alpha_\pi (r^\pi - g^*\tau^{\pi,*}) = \frac{r \cdot x^* - g^*\tau^* \cdot x^*}{\Lambda^*} = 0.$$

Hence

$$\alpha_\pi > 0 \implies r^\pi/\tau^{\pi,*} = g^*. \tag{49}$$

Set $\chi^*_\pi = \Lambda^*\alpha_\pi$ and $\nu^*_\pi = \chi^*_\pi \tau^{\pi,*}$. The decomposition gives $\mathbf{R}\chi^* = x^*$, hence $\tau^\pi(\chi^*) = \tau^{\pi,*}$. The mass constraint gives

$$\sum_\pi \nu^*_\pi = \tau^* \cdot x^* = m.$$

By construction, $\nu^*_\pi = \chi^*_\pi \tau^\pi(\chi^*)$. Uniqueness in (7) therefore implies $\chi^* = \bar{\chi}^{\nu^*}$, and (49) gives $\nu^* \in \mathcal{E}^*_\Pi$. Moreover, $\bar{x}^{\nu^*} = \mathbf{R}\chi^* = x^*$, so

$$\bar{\mu}^{\nu^*} = \tau^* \odot x^* = \mu^*.$$

Consistency of the original equilibrium also gives (11) for π^* at every positive-throughput state. This proves Part (i).

(ii) For every $\nu^* \in \mathcal{E}^*_\Pi$, Theorem 5 gives $\nabla G(\bar{x}^{\nu^*}) = \tau^*$. Consequently, each policy's mean duration is constant on $\mathcal{E}^*_\Pi$:

$$\tau^\pi(\bar{\chi}^{\nu^*}) = \rho^\pi \cdot \tau^* = \tau^{\pi,*}.$$

Little's law, $\nu^*_\pi = \bar{\chi}^{\nu^*}_\pi \tau^{\pi,*}$, gives $\bar{\chi}^{\nu^*}_\pi = \nu^*_\pi/\tau^{\pi,*}$. The payoff definition gives $H_\pi(\nu^*) = r^\pi/\tau^{\pi,*} = H^*_\pi$. The completion rates are $\lambda_i(\bar{\mu}^{\nu^*}) = 1/\tau^*_i$, and LP optimality gives $r \cdot \bar{x}^{\nu^*}/m = g^*$. Lemma 22 gives $\max_\pi H^*_\pi = g^*$.

(iii) Compactness of $\mathcal{E}^*_\Pi$ supplies a nearest $\nu^*(t) \in \mathcal{E}^*_\Pi$, with $\|\nu(t) - \nu^*(t)\| = \mathrm{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$. Uniform continuity of the population map gives

$$\mathrm{dist}(\bar{\mu}^{\nu(t)}, \mathcal{E}^*_\mathcal{I}) \le \|\bar{\mu}^{\nu(t)} - \bar{\mu}^{\nu^*(t)}\| \longrightarrow 0.$$

Similarly, uniform continuity of H gives

$$\|H(\nu(t)) - H^*\| = \|H(\nu(t)) - H(\nu^*(t))\| \longrightarrow 0.$$

The identical comparison applies to $\lambda(\bar{\mu}^{\nu(t)})$ and $r \cdot \bar{x}^{\nu(t)}/m$, using their constant values in Part (ii). □

Relation between the policy and equilibrium-program gaps. The policy potential gap $\Delta(\nu)$ also bounds the objective gap of the feasible completion-rate vector $\bar{x}^\nu$ in (3). Write $\Psi_{\mathrm{EG}}(x) = m\log(r \cdot x) - G(x)$, with optimal value Ψ^*_{EG}. On $\mathcal{E}^*_\Pi$,

$$\Phi^* = m\log g^* - G(\bar{x}^\nu) = \Psi^*_{\mathrm{EG}} - m\log m.$$

For arbitrary ν, use $r \cdot \bar{x}^\nu = \sum_\pi \nu_\pi H_\pi(\nu)$ in (37) to obtain

$$\Delta(\nu) = \Psi^*_{\mathrm{EG}} - \Psi_{\mathrm{EG}}(\bar{x}^\nu) + m\log\left(\frac{1}{m}\sum_\pi \nu_\pi H_\pi(\nu)\right) - \sum_\pi \nu_\pi \log H_\pi(\nu). \tag{50}$$

Feasibility of $\bar{x}^\nu$ and Jensen's inequality imply $0 \le \Psi^*_{\mathrm{EG}} - \Psi_{\mathrm{EG}}(\bar{x}^\nu) \le \Delta(\nu)$.

D Proofs and Technical Details for Section 5

We first establish existence for the state–policy dynamics with exogenous revision. Holding the policy distribution fixed then reduces physical convergence to transient decay and a recurrent Lyapunov argument, as in Appendix B. We then apply trajectory compactness to identify the cluster points of omega-limit populations as revisions become rare. Together with policy convergence, this proves Theorem 11. Appendix D.5 extends the analysis to exogenous BNN and replicator revision.

D.1 Existence and the policy marginal equation

Proposition 28. *Under Assumptions 1 and 4, for every $\kappa \geq 0$, the state–policy differential inclusion (14) admits at least one global absolutely continuous solution from every initial point of $\mathcal{M}_m(\mathcal{S} \times \Pi)$. The simplex is forward invariant, total mass is conserved, and every solution has policy marginal*

$$\dot{\nu} \in \kappa(\Gamma(\nu) - \nu)$$

almost everywhere. If a single-valued locally Lipschitz selection $\gamma(\nu) \in \Gamma(\nu)$ is fixed, the state–policy trajectory is unique.

Proof. For $\gamma \in \Gamma(\nu)$, an Euler step is

$$\eta^+_{s,\pi} = [1 - \Delta t(\lambda_{s,\pi(s)}(\mu) + \kappa)]\eta_{s,\pi} + \Delta t \sum_{s'} X_{s',\pi} P^{\pi(s')}_{s's} + \Delta t \kappa \frac{\gamma_\pi}{m} m_s(\eta).$$

For $0 < \Delta t \leq (\overline{\lambda} + \kappa)^{-1}$, all terms are nonnegative, so $\eta^+ \geq 0$. Normalization of γ gives

$$\sum_{s,\pi} \eta^+_{s,\pi} = m + \Delta t \kappa \left(\frac{\sum_\pi \gamma_\pi}{m} \sum_s m_s - m \right) = m.$$

The map $\eta \mapsto \nu = (\sum_s \eta_{s,\pi})_{\pi \in \Pi}$ in (12) is linear, and the drift in (14) is continuous in η and affine in γ. Lemma 24 therefore gives a global solution with a common measurable selection $\gamma(t) \in \Gamma(\nu(t))$.

For every solution, summing (14) over states gives

$$\dot{\nu}_\pi = \sum_{s'} X_{s',\pi} \left(\sum_s P^{\pi(s')}_{s's} - 1 \right) + \kappa \left(\frac{\gamma_\pi}{m} \sum_s m_s - \nu_\pi \right) = \kappa(\gamma_\pi - \nu_\pi).$$

Summing over policies gives $\sum_\pi \dot{\nu}_\pi = 0$. For $\kappa > 0$, the best-response vector used in (14) is $\gamma = \nu + \dot{\nu}/\kappa$ almost everywhere, hence measurable. For $\kappa = 0$, the right-hand side of (14) is independent of γ. For each component, the inflow rate is $\sum_{s'} X_{s',\pi}(\eta) P^{\pi(s')}_{s's} + \kappa \gamma_\pi m_s(\eta)/m \geq 0$, and the outflow rate is $[\lambda_{s,\pi(s)}(\mu) + \kappa]\eta_{s,\pi}$, by (13). Thus $\dot{\eta}_{s,\pi} + [\lambda_{s,\pi(s)}(\mu) + \kappa]\eta_{s,\pi} \geq 0$ almost everywhere. Multiplying by a positive exponential and applying the product rule gives

$$\frac{d}{dt} \left[\eta_{s,\pi}(t) \exp\left(\int_0^t [\lambda_{s,\pi(s)}(\mu(t')) + \kappa]\, dt' \right) \right] \geq 0.$$

Integrating from 0 to t and dividing by the exponential yields

$$\eta_{s,\pi}(t) \geq \eta_{s,\pi}(0) \exp\left(- \int_0^t [\lambda_{s,\pi(s)}(\mu(t')) + \kappa]\, dt' \right) \geq 0.$$

For uniqueness, fix a locally Lipschitz selection $\nu \mapsto \gamma(\nu) \in \Gamma(\nu)$. The maps $\eta \mapsto \mu$, $\eta \mapsto \nu$, and $\eta \mapsto m_s(\eta)$ are linear, while $\mu \mapsto \lambda(\mu)$ is locally Lipschitz by Proposition 2. Hence both $X_{s,\pi}(\eta) = \eta_{s,\pi} \lambda_{s,\pi(s)}(\mu(\eta))$ and $\gamma_\pi(\nu(\eta)) m_s(\eta)$ are locally Lipschitz in η. Every term on the right-hand side of (14) is therefore locally Lipschitz, so ODE uniqueness applies (Teschl, 2012, Theorem 2.2). □

D.2 Proof of Proposition 10: convergence of the frozen state–policy dynamics

We establish transient decay (57) and strict Lyapunov dissipation (60) for the frozen state–policy dynamics (51). Together they imply that every complete solution is stationary, which yields the proposition and will also identify rare-revision limits.

Fix $\nu \in \mathcal{M}_m(\Pi)$. Setting $\kappa = 0$ in (14) gives

$$\dot{\eta}_{s,\pi}(t) = \sum_{s' \in \mathcal{S}} X_{s',\pi}(\eta(t)) P^{\pi(s')}_{s's} - X_{s,\pi}(\eta(t)). \tag{51}$$

Denote its vector field by F_{phys}. It preserves every policy mass. Write

$$\mathcal{E}_\nu := \left\{ \eta \in \mathcal{M}_m(\mathcal{S} \times \Pi) : \sum_s \eta_{s,\pi} = \nu_\pi \ \forall \pi \right\}. \tag{52}$$

Thus $\mathcal{E}_\nu$ consists of population vectors with the fixed policy masses ν_π. The frozen field is locally Lipschitz by Proposition 2; its inflow terms are nonnegative, and summing over s gives zero for each policy. Hence $\mathcal{E}_\nu$ is forward invariant, and compactness gives a unique global solution for each initial condition in $\mathcal{E}_\nu$ (Teschl, 2012, Theorem 2.2 and Corollary 2.15), with continuous dependence on its initial condition.

For every $\pi \in \Pi$, let $\mathcal{S}^\pi_{\mathrm{rec}}$ and $\mathcal{S}^\pi_{\mathrm{tr}}$ be its recurrent and transient state sets. The recurrent state–policy pairs form the set $\mathcal{J}_{\mathrm{rec}} := \{(s,\pi) : s \in \mathcal{S}^\pi_{\mathrm{rec}}\}$. For each state–policy pair (s,π), the stationary completion rate is

$$\bar{X}^\nu_{s,\pi} := \bar{\chi}^\nu_\pi q^\pi_s, \tag{53}$$

so $\bar{X}^\nu_{s,\pi} > 0$ on $\mathcal{S}^\pi_{\mathrm{rec}}$ when $\nu_\pi > 0$ and $\bar{X}^\nu_{s,\pi} = 0$ on $\mathcal{S}^\pi_{\mathrm{tr}}$. Summing over policies that prescribe action a at state s gives $\bar{x}^\nu_{sa} = \sum_{\pi:\pi(s)=a} \bar{X}^\nu_{s,\pi}$. These state–action rates form the aggregate vector $\bar{x}^\nu = (\bar{x}^\nu_{sa})_{(s,a)\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}_+$ in (8). The policy mass balances give $\bar{\eta}^\nu$ from (16). In particular, $x(\bar{\mu}^\nu) = \bar{x}^\nu$, so

$$X_{s,\pi}(\bar{\eta}^\nu) = \frac{\bar{\eta}^\nu_{s,\pi}}{\partial_{s,\pi(s)} G(\bar{x}^\nu)} = \bar{X}^\nu_{s,\pi}.$$

Invariance of q^π then gives state balance:

$$\sum_{s'} \bar{X}^\nu_{s',\pi} P^{\pi(s')}_{s's} = \bar{\chi}^\nu_\pi q^\pi_s = \bar{X}^\nu_{s,\pi}.$$

Thus $\bar{\eta}^\nu$ is stationary for (51).

The state–policy analogue of (2) is

$$\widetilde{f}^*(\eta) := \max_{X' \in \mathbb{R}^{\mathcal{S}\times\Pi}_+} \left\{ \sum_{s,\pi} \eta_{s,\pi} \log X'_{s,\pi} - G(BX') \right\}, \tag{54}$$

where $B : \mathbb{R}^{\mathcal{S}\times\Pi} \to \mathbb{R}^{\mathcal{I}}$ is the linear map aggregating state–policy rates into state–action rates, i.e., $B_{(s,a),(s',\pi)} = \mathbf{1}_{\{s=s',\, a=\pi(s)\}}$, and satisfies $(BX')_{sa} = \sum_{\pi:\pi(s)=a} X'_{s,\pi}$. The same map aggregates populations: $\mu(\eta) = B\eta$, and $x(\mu(\eta)) = BX(\eta)$. The extended-value convention in (2) applies. On the open convex neighborhood $B^{-1}(U) \supset \mathbb{R}^{\mathcal{S}\times\Pi}_+$, the function $X' \mapsto G(BX')$ is C^2 and convex, since

$$\nabla^2_{X'}[G(BX')] = B^\top \nabla^2 G(BX') B \succeq 0.$$

Its partial derivatives are positive on $X' \geq 0$:

$$\frac{\partial}{\partial X'_{s,\pi}} G(BX') = \partial_{s,\pi(s)} G(BX') > 0 \quad (X' \geq 0).$$

Proposition 2 gives a unique optimizer X' of (54), with the same support as η, and a local C^1 extension of the optimizer map. Little's law at this optimizer reads

$$\eta_{s,\pi} = X'_{s,\pi} \partial_{s,\pi(s)} G(BX'). \tag{55}$$

Summing (55) over policies with $\pi(s) = a$ gives $\mu_{sa} = (BX')_{sa} \partial_{sa} G(BX')$, hence $BX' = x(\mu)$ by uniqueness. Substituting back gives $X' = X(\eta)$ as defined in (13). Thus (54) has optimizer $X(\eta)$ and optimal value $\widetilde{f}^*(\eta)$.

The next lemma proves exponential decay of the mass in transient states for each policy.

Lemma 29. *Under Assumptions 1 and 4, for $\mathcal{S}^\pi_{\mathrm{tr}} \neq \varnothing$, let*

$$P^\pi_{\mathrm{tr}} := \left(P^{\pi(s)}_{ss'}\right)_{s,s'\in\mathcal{S}^\pi_{\mathrm{tr}}}, \qquad w^\pi := (\mathrm{I} - P^\pi_{\mathrm{tr}})^{-1}\mathbf{1}.$$

Define

$$W_\pi(\eta) := \sum_{s\in\mathcal{S}^\pi_{\mathrm{tr}}} w^\pi_s \eta_{s,\pi}, \qquad W(\eta) := \sum_\pi W_\pi(\eta),$$

with $W_\pi \equiv 0$ when $\mathcal{S}^\pi_{\mathrm{tr}} = \varnothing$. Then $w^\pi_s \geq 1$ for every $s \in \mathcal{S}^\pi_{\mathrm{tr}}$, and along the frozen state–policy dynamics (51)

$$\frac{d}{dt} W_\pi(\eta(t)) = -\sum_{s\in\mathcal{S}^\pi_{\mathrm{tr}}} X_{s,\pi}(\eta(t)). \tag{56}$$

Consequently there is $c_{\mathrm{tr}} > 0$ such that

$$\dot{W} \le -c_{\mathrm{tr}} W. \tag{57}$$

In particular, every complete solution $\zeta : \mathbb{R} \to \mathcal{E}_\nu$ of (51) *satisfies $W(\zeta(t)) = 0$ for all $t \in \mathbb{R}$.*

Proof. For each policy with $\mathcal{S}^\pi_{\mathrm{tr}} \neq \varnothing$, the calculation in the proof of Lemma 19 applies with P^π_{tr}, $\eta_{s,\pi}$, and $X_{s,\pi}$ in place of p^π_{tr}, μ_i, and x_i, respectively. It gives $w^\pi \ge \mathbf{1}$ and $(\mathrm{I} - P^\pi_{\mathrm{tr}})w^\pi = \mathbf{1}$. Write $\eta_{\mathcal{S}^\pi_{\mathrm{tr}},\pi} := (\eta_{s,\pi})_{s\in\mathcal{S}^\pi_{\mathrm{tr}}}$ and define $X_{\mathcal{S}^\pi_{\mathrm{tr}},\pi}$ similarly. Since $\mathcal{S}^\pi_{\mathrm{rec}}$ is closed, $\dot{\eta}_{\mathcal{S}^\pi_{\mathrm{tr}},\pi} = ((P^\pi_{\mathrm{tr}})^\top - \mathrm{I})X_{\mathcal{S}^\pi_{\mathrm{tr}},\pi}$. Hence

$$\dot{W}_\pi = (w^\pi)^\top((P^\pi_{\mathrm{tr}})^\top - \mathrm{I})X_{\mathcal{S}^\pi_{\mathrm{tr}},\pi} = -\sum_{s\in\mathcal{S}^\pi_{\mathrm{tr}}} X_{s,\pi},$$

which is (56). To obtain a common decay bound for the finite policy family, set $c_{\mathrm{tr}} := \underline{\lambda}/\max(\{1\} \cup \{w^\pi_s : \pi \in \Pi,\ s \in \mathcal{S}^\pi_{\mathrm{tr}}\}) > 0$. Since $X_{s,\pi} = \eta_{s,\pi}\lambda_{s,\pi(s)}(\mu) \ge \underline{\lambda}\eta_{s,\pi}$, summing over policies gives

$$\dot{W} \le -\underline{\lambda}\sum_\pi \sum_{s\in\mathcal{S}^\pi_{\mathrm{tr}}} \eta_{s,\pi} \le -c_{\mathrm{tr}} W.$$

Empty transient blocks contribute zero, so this also covers policies without transient states. Since $\mathcal{E}_\nu$ is compact, the argument for complete solutions in Lemma 19 applies to (57) and gives $W(\zeta(t)) = 0$ for all $t \in \mathbb{R}$. □

If $\nu_\pi = 0$, nonnegativity and $\sum_s \eta_{s,\pi} = \nu_\pi$ on $\mathcal{E}_\nu$ imply $\eta_{s,\pi} = X_{s,\pi}(\eta) = 0$ for every s. The policy mass balances and (53) also give $\bar{\chi}^\nu_\pi = \bar{X}^\nu_{s,\pi} = 0$. The logarithmic expression below in (59) therefore omits policies with $\nu_\pi = 0$.

Define the recurrent face of (52) by

$$\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}} := \{\eta \in \mathcal{E}_\nu : \eta_{s,\pi} = 0 \text{ for every } \pi \in \Pi \text{ and } s \in \mathcal{S}^\pi_{\mathrm{tr}}\}. \tag{58}$$

Its relative interior, denoted by $\mathrm{relint}(\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}})$, consists of the vectors in $\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}}$ with $\eta_{s,\pi} > 0$ whenever $\nu_\pi > 0$ and $s \in \mathcal{S}^\pi_{\mathrm{rec}}$. All other components are zero. For every $\eta \in \mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}}$, define

$$\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}(\eta) := \widetilde{f}^*(\eta) - \sum_{\pi:\nu_\pi>0}\sum_{s\in\mathcal{S}^\pi_{\mathrm{rec}}} \eta_{s,\pi}\log \bar{X}^\nu_{s,\pi} + G(\bar{x}^\nu). \tag{59}$$

The logarithms in the sum are finite because $\bar{X}^\nu_{s,\pi} > 0$ whenever $\nu_\pi > 0$ and $s \in \mathcal{S}^\pi_{\mathrm{rec}}$. The next lemma is the state–policy counterpart of Proposition 6.

Lemma 30. *Under Assumptions 1 and 4,*

(i) $\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}(\eta) \ge 0$ for every $\eta \in \mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}}$, with equality if and only if $\eta = \bar{\eta}^\nu$.

(ii) Along every solution of (51) *in* $\mathrm{relint}(\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}})$*,*

$$\frac{d}{dt}\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}(\eta(t)) = \sum_{\pi:\nu_\pi>0}\sum_{s,s'\in\mathcal{S}^\pi_{\mathrm{rec}}} X_{s,\pi}P^{\pi(s)}_{ss'}\log\frac{X_{s',\pi}/\bar{X}^\nu_{s',\pi}}{X_{s,\pi}/\bar{X}^\nu_{s,\pi}} \le 0. \tag{60}$$

The derivative is zero if and only if $\eta = \bar{\eta}^\nu$.

Proof. (i) As in the proof of Proposition 6(i), evaluate the objective in (54) at the stationary rates $\bar{X}^\nu$:

$$\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}(\eta) = \widetilde{f}^*(\eta) - \left(\sum_{\pi:\nu_\pi>0}\sum_{s\in\mathcal{S}^\pi_{\mathrm{rec}}} \eta_{s,\pi}\log\bar{X}^\nu_{s,\pi} - G(\bar{x}^\nu)\right) \ge 0.$$

Uniqueness of the optimizer in (54) and Little's law (55) give

$$\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}(\eta) = 0 \iff X(\eta) = \bar{X}^\nu \iff \eta = \bar{\eta}^\nu.$$

This argument also applies at the boundary of $\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}}$.

(ii) Write $X = X(\eta)$ and $x = x(\mu(\eta)) = BX$. State–policy sums below are restricted to policies with $\nu_\pi > 0$ and states in $\mathcal{S}^\pi_{\mathrm{rec}}$, where $X_{s,\pi}, \bar{X}^\nu_{s,\pi} > 0$. The remaining state–policy masses and rates

stay zero on the recurrent face $\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}}$. By Proposition 2, applied to $X' \mapsto G(BX')$, the optimizer map $\eta \mapsto X(\eta)$ is C^1. Since $x_{sa} = \sum_{\pi:\pi(s)=a} X_{s,\pi}$, the chain rule gives

$$dG(x) = \sum_{s,a} \partial_{sa} G(x)\, dx_{sa} = \sum_{s,a} \partial_{sa} G(x) \sum_{\pi:\pi(s)=a} dX_{s,\pi} = \sum_{s,\pi} \partial_{s,\pi(s)} G(x)\, dX_{s,\pi}.$$

Evaluating (54) at $X(\eta)$, applying the product rule, and then using Little's law (55), we obtain

$$\begin{aligned}
d\widetilde{f}^* &= d\left(\sum_{s,\pi} \eta_{s,\pi} \log X_{s,\pi} - G(x)\right) \\
&= \sum_{s,\pi} \left(\log X_{s,\pi}\, d\eta_{s,\pi} + \eta_{s,\pi}\, d(\log X_{s,\pi})\right) - dG(x) \\
&= \sum_{s,\pi} \log X_{s,\pi}\, d\eta_{s,\pi} + \sum_{s,\pi} \left(\frac{\eta_{s,\pi}}{X_{s,\pi}} - \partial_{s,\pi(s)} G(x)\right) dX_{s,\pi} \\
&= \sum_{s,\pi} \log X_{s,\pi}\, d\eta_{s,\pi} + \sum_{s,\pi} \left(\frac{X_{s,\pi} \partial_{s,\pi(s)} G(x)}{X_{s,\pi}} - \partial_{s,\pi(s)} G(x)\right) dX_{s,\pi} \\
&= \sum_{s,\pi} \log X_{s,\pi}\, d\eta_{s,\pi}.
\end{aligned}$$

The differential identity above therefore makes $\widetilde{f}^*$, and hence $\mathcal{L}_\nu^{\mathcal{J}_{\mathrm{rec}}}$, C^2 on $\mathrm{relint}(\mathcal{E}_\nu^{\mathcal{J}_{\mathrm{rec}}})$.

Along the frozen dynamics, ν is fixed, so $\bar{X}^\nu$ and $\bar{x}^\nu$ are constant. Each recurrent class is closed, and its stationary rates satisfy state balance; thus, on each such class,

$$\sum_{s'} P_{ss'}^{\pi(s)} = 1, \qquad \sum_{s} \bar{X}_{s,\pi}^\nu P_{ss'}^{\pi(s)} = \bar{X}_{s',\pi}^\nu.$$

Differentiating (59), substituting (51), exchanging s and s' in the inflow sum, and expanding the outflow using the row-sum identity gives

$$\begin{aligned}
\dot{\mathcal{L}}_\nu^{\mathcal{J}_{\mathrm{rec}}} &= \sum_{s,\pi} \dot{\eta}_{s,\pi} \big(\log X_{s,\pi} - \log \bar{X}_{s,\pi}^\nu\big) \\
&= \sum_{s,\pi} \left(\sum_{s'} X_{s',\pi} P_{s's}^{\pi(s')} - X_{s,\pi}\right) \log \frac{X_{s,\pi}}{\bar{X}_{s,\pi}^\nu} \\
&= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \log \frac{X_{s',\pi}}{\bar{X}_{s',\pi}^\nu} - \sum_{s,\pi} X_{s,\pi} \log \frac{X_{s,\pi}}{\bar{X}_{s,\pi}^\nu} \\
&= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \log \frac{X_{s',\pi}}{\bar{X}_{s',\pi}^\nu} - \sum_{s,\pi} X_{s,\pi} \left(\sum_{s'} P_{ss'}^{\pi(s)}\right) \log \frac{X_{s,\pi}}{\bar{X}_{s,\pi}^\nu} \\
&= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \left(\log \frac{X_{s',\pi}}{\bar{X}_{s',\pi}^\nu} - \log \frac{X_{s,\pi}}{\bar{X}_{s,\pi}^\nu}\right) \\
&= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \log \frac{X_{s',\pi}/\bar{X}_{s',\pi}^\nu}{X_{s,\pi}/\bar{X}_{s,\pi}^\nu}.
\end{aligned}$$

The row-sum and stationary-balance identities also give the cancellation

$$\begin{aligned}
&\sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \left(\frac{X_{s',\pi}/\bar{X}^{\nu}_{s',\pi}}{X_{s,\pi}/\bar{X}^{\nu}_{s,\pi}} - 1 \right) \\
&= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \frac{X_{s',\pi}\bar{X}^{\nu}_{s,\pi}}{\bar{X}^{\nu}_{s',\pi} X_{s,\pi}} - \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \\
&= \sum_{\pi,s,s'} \bar{X}^{\nu}_{s,\pi} P_{ss'}^{\pi(s)} \frac{X_{s',\pi}}{\bar{X}^{\nu}_{s',\pi}} - \sum_{s,\pi} X_{s,\pi} \sum_{s'} P_{ss'}^{\pi(s)} \\
&= \sum_{s',\pi} \frac{X_{s',\pi}}{\bar{X}^{\nu}_{s',\pi}} \left(\sum_{s} \bar{X}^{\nu}_{s,\pi} P_{ss'}^{\pi(s)} \right) - \sum_{s,\pi} X_{s,\pi} \\
&= \sum_{s',\pi} \frac{X_{s',\pi}}{\bar{X}^{\nu}_{s',\pi}} \bar{X}^{\nu}_{s',\pi} - \sum_{s,\pi} X_{s,\pi} = \sum_{s',\pi} X_{s',\pi} - \sum_{s,\pi} X_{s,\pi} = 0.
\end{aligned}$$

Subtracting this zero sum from the logarithmic sum yields

$$\begin{aligned}
\dot{\mathcal{L}}_{\nu}^{\mathcal{J}_{\mathrm{rec}}} &= \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \left[\log \frac{X_{s',\pi}/\bar{X}^{\nu}_{s',\pi}}{X_{s,\pi}/\bar{X}^{\nu}_{s,\pi}} - \left(\frac{X_{s',\pi}/\bar{X}^{\nu}_{s',\pi}}{X_{s,\pi}/\bar{X}^{\nu}_{s,\pi}} - 1 \right) \right] \\
&= - \sum_{\pi,s,s'} X_{s,\pi} P_{ss'}^{\pi(s)} \left(\frac{X_{s',\pi}/\bar{X}^{\nu}_{s',\pi}}{X_{s,\pi}/\bar{X}^{\nu}_{s,\pi}} - 1 - \log \frac{X_{s',\pi}/\bar{X}^{\nu}_{s',\pi}}{X_{s,\pi}/\bar{X}^{\nu}_{s,\pi}} \right) \leq 0.
\end{aligned}$$

This is (35) applied to each stationary recurrent block. The inequality follows from $z - 1 - \log z \geq 0$ for $z > 0$, with equality exactly at $z = 1$, and the nonnegative weights $X_{s,\pi} P_{ss'}^{\pi(s)}$. Its equality condition and irreducibility give

$$\dot{\mathcal{L}}_{\nu}^{\mathcal{J}_{\mathrm{rec}}} = 0 \iff X_{s,\pi} = k_\pi \bar{X}^{\nu}_{s,\pi} \quad (k_\pi > 0,\ s \in \mathcal{S}^{\pi}_{\mathrm{rec}},\ \nu_\pi > 0).$$

There is one scaling constant k_π for each policy with $\nu_\pi > 0$, corresponding to the single constant k in the proof of Proposition 6(ii). Summing over states gives $\chi_\pi(\eta) = k_\pi \bar{\chi}^{\nu}_{\pi}$ when $\nu_\pi > 0$, and $\chi_\pi(\eta) = 0$ otherwise. Thus $(\mathbf{R}\chi(\eta))_{sa} = \sum_{\pi:\pi(s)=a} X_{s,\pi} = x_{sa}(\mu(\eta))$. The policy mass equations therefore read

$$\nu_\pi = \sum_s \eta_{s,\pi} = \chi_\pi(\eta) \sum_s q_s^\pi \partial_{s,\pi(s)} G(\mathbf{R}\chi(\eta)) = \chi_\pi(\eta) \tau^\pi(\chi(\eta)).$$

Both $\chi(\eta)$ and $\bar{\chi}^{\nu}$ satisfy the mass balances $\nu_\pi = \chi_\pi \tau^\pi(\chi)$. Uniqueness in (7) gives $\chi(\eta) = \bar{\chi}^{\nu}$. Hence $k_\pi = 1$ for every policy with $\nu_\pi > 0$, so $X_{s,\pi} = \bar{X}^{\nu}_{s,\pi}$. Little's law (55) then gives $\eta = \bar{\eta}^{\nu}$. Conversely, at $\bar{\eta}^{\nu}$ we have $X_{s,\pi} = \bar{X}^{\nu}_{s,\pi}$, so (60) vanishes. □

Corollary 31. *Under Assumptions 1 and 4, every complete solution $\zeta : \mathbb{R} \to \mathcal{E}_\nu$ of* (51) *is constant and equals $\bar{\eta}^{\nu}$.*

Proof. We follow the complete-trajectory argument in Appendix B.4. Lemma 29 gives $W(\zeta(t)) = 0$, hence $\zeta(t) \in \mathcal{E}_{\nu}^{\mathcal{J}_{\mathrm{rec}}}$ for every $t \in \mathbb{R}$. For each policy π with $\nu_\pi > 0$, restrict (51) to $\mathcal{S}^{\pi}_{\mathrm{rec}}$. Its transition matrix $P = (P_{ss'}^{\pi(s)})_{s,s' \in \mathcal{S}^{\pi}_{\mathrm{rec}}}$ is irreducible and stochastic, its mass is ν_π, and its completion rates lie in $[\underline{\lambda}, \overline{\lambda}]$. Lemma 20, applied to the chain on $\mathcal{S}^{\pi}_{\mathrm{rec}}$ with mass ν_π, therefore gives

$$\zeta_{s,\pi}(t) \geq c_{\mathcal{S}^{\pi}_{\mathrm{rec}}} \nu_\pi > 0 \quad (\nu_\pi > 0,\ s \in \mathcal{S}^{\pi}_{\mathrm{rec}},\ t \in \mathbb{R}).$$

Since there are finitely many positive-mass policies, these bounds put the orbit closure in a compact subset of $\mathrm{relint}(\mathcal{E}_{\nu}^{\mathcal{J}_{\mathrm{rec}}})$. On this set, Lemma 30 gives $\mathcal{L}_{\nu}^{\mathcal{J}_{\mathrm{rec}}} \geq 0$ and $\dot{\mathcal{L}}_{\nu}^{\mathcal{J}_{\mathrm{rec}}} \leq 0$, with zero derivative only at $\eta = \bar{\eta}^{\nu}$. Lemma 18 therefore implies $\zeta \equiv \bar{\eta}^{\nu}$. □

Proof of Proposition 10. Existence and uniqueness were established after (52), and compactness of $\mathcal{E}_\nu$ makes $\omega(\eta)$ nonempty. We use the same time-shift construction as in the proof of Theorem 7.

For $\eta^\infty \in \omega(\eta)$, choose $t_n \geq n$ with $\eta(t_n) \to \eta^\infty$, and set $\eta_n(t) = \eta(t_n + t)$, $|t| \leq n$. The shifted trajectories remain in $\mathcal{E}_\nu$, so $\|\eta_n(t)\| \leq m$. Their derivatives satisfy

$$\|\dot{\eta}_n(t)\| \leq \|\dot{\eta}_n(t)\|_1 \leq 2 \sum_{s,\pi} X_{s,\pi}(\eta_n(t)) \leq 2m\overline{\lambda}. \tag{61}$$

Lemma 17 gives a subsequence η_{n_k} and a complete frozen solution $\zeta : \mathbb{R} \to \mathcal{E}_\nu$ such that

$$\sup_{|t| \leq T} \|\eta_{n_k}(t) - \zeta(t)\| \longrightarrow 0 \qquad \text{for every } T < \infty.$$

In particular, $\zeta(0) = \eta^\infty$. Corollary 31 yields $\eta^\infty = \zeta(0) = \bar{\eta}^\nu$. Since η^∞ was arbitrary, $\omega(\eta) = \{\bar{\eta}^\nu\}$; compactness of $\mathcal{E}_\nu$ then gives $\eta(t) \to \bar{\eta}^\nu$. A stationary solution is complete, so the same corollary gives uniqueness of the stationary population. □

D.3 Geometry of the stationary target set

Under Assumptions 1 and 4, the set $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ from Section 5 is nonempty, compact, and convex. Indeed, the policy equilibrium set $\mathcal{E}^*_\Pi$ is nonempty and compact by Theorem 8, and $\nu \mapsto \bar{\eta}^\nu$ is continuous by the explicit formula (16) and Lemma 23. Thus $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ is nonempty and compact. For every $\nu^* \in \mathcal{E}^*_\Pi$, Theorems 5(ii) and 9(i) give $\nabla G(\bar{x}^{\nu^*}) = \tau^*$, with the same τ^* throughout $\mathcal{E}^*_\Pi$. Since q^π depends only on the fixed policy and transition matrix, $\tau^{\pi,*} = \sum_s q^\pi_s \tau^*_{s,\pi(s)} > 0$ is also independent of ν^*. Theorem 9(ii) and (16) therefore give

$$\bar{\eta}^{\nu^*}_{s,\pi} = \frac{\nu^*_\pi}{\tau^{\pi,*}} q^\pi_s \tau^*_{s,\pi(s)}.$$

Thus $\nu \mapsto \bar{\eta}^\nu$ agrees on $\mathcal{E}^*_\Pi$ with a fixed linear map. Since $\mathcal{E}^*_\Pi$ is convex, so is $\mathcal{E}^*_{\mathcal{S}\times\Pi}$. The marginal map (12) recovers ν from $\bar{\eta}^\nu$, so this linear correspondence is injective; it does not imply uniqueness of the population equilibrium or of its policy decomposition.

D.4 Proof of Theorem 11: rare exogenous revisions

The policy marginal converges by Theorem 8. Trajectory compactness and stationarity of complete frozen solutions then identify the rare-revision limit of the state–policy population.

Proof of Theorem 11. Global existence is Proposition 28. (i) The marginal equation is $\dot{\nu}^\kappa = \kappa(\gamma - \nu^\kappa)$. Theorem 8 gives

$$\frac{d}{dt}\Phi(\nu^\kappa(t)) = \kappa M_{\mathrm{BR}}(\nu^\kappa(t)) \geq 0. \tag{62}$$

With slow time $\widetilde{t} = \kappa t$, applying its potential-gap bound to the rescaled trajectory $\widetilde{t} \mapsto \nu^\kappa(\widetilde{t}/\kappa)$ gives

$$\Delta(\nu^\kappa(t)) \leq e^{-\kappa t}\Delta(\nu^\kappa(0)).$$

Its best-response gap bound similarly gives

$$M_{\mathrm{BR}}(\nu^\kappa(t)) \leq e^{-\kappa t} M_{\mathrm{BR}}(\nu^\kappa(0)).$$

Thus $\mathrm{dist}(\nu^\kappa(t), \mathcal{E}^*_\Pi) \to 0$.

(ii) Let $\kappa_n \downarrow 0$, let η^{κ_n} be arbitrary solutions of (14) from initial populations in $\mathcal{M}_m(\mathcal{S} \times \Pi)$, and choose $\eta^\infty_n \in \omega(\eta^{\kappa_n})$. Consider any convergent subsequence $\eta^\infty_n \to \eta^*$, relabeled by n. By Part (i), the policy marginal of each η^∞_n belongs to $\mathcal{E}^*_\Pi$. Since the marginal map is continuous and $\mathcal{E}^*_\Pi$ is closed, the marginal ν^* of η^* also belongs to $\mathcal{E}^*_\Pi$.

For each n, the definition of an omega-limit point gives a time $t_n \geq n$ such that $\|\eta^{\kappa_n}(t_n) - \eta^\infty_n\| \leq n^{-1}$. Set

$$\zeta_n(t) := \eta^{\kappa_n}(t_n + t), \qquad |t| \leq n.$$

These trajectories lie in the compact state–policy simplex and satisfy $\zeta_n(0) \to \eta^*$. In the revision term of (14), the inflow vector $(\gamma_\pi m_s(\eta)/m)_{s,\pi}$ and the outflow vector η are nonnegative and each sums to m. Consequently,

$$\|\dot{\zeta}_n(t) - F_{\mathrm{phys}}(\zeta_n(t))\|_1 \leq 2m\kappa_n \quad \text{a.e.}$$

Then it follows that

$$\|\dot{\zeta}_n(t)\|_1 \le 2m(\overline{\lambda} + \kappa_1) \quad \text{a.e.},$$

where the frozen-field bound follows from (61). For every fixed $T < \infty$ and $n \ge T$, integration gives

$$\sup_{a,b\in[-T,T]} \left\| \zeta_n(b) - \zeta_n(a) - \int_a^b F_{\mathrm{phys}}(\zeta_n(t'))\, dt' \right\| \le 4mT\kappa_n \longrightarrow 0.$$

Apply Lemma 17 with $I = \mathbb{R}$, $K = \mathcal{M}_m(\mathcal{S} \times \Pi)$, $T_n = n$, and $f = F_{\mathrm{phys}}$. Part (i) gives a subsequence $(\zeta_{n_k})_{k\ge1}$ and a Lipschitz map $\zeta : \mathbb{R} \to \mathcal{M}_m(\mathcal{S} \times \Pi)$ such that

$$\sup_{|t|\le T} \|\zeta_{n_k}(t) - \zeta(t)\| \longrightarrow 0 \quad (k \to \infty) \qquad \text{for every } T > 0.$$

The residual estimate above and Part (ii) give

$$\zeta(b) - \zeta(a) = \int_a^b F_{\mathrm{phys}}(\zeta(t'))\, dt' \qquad \text{for all } a, b \in \mathbb{R}.$$

Thus ζ is a complete solution of (51), and $\zeta(0) = \lim_{k\to\infty} \zeta_{n_k}(0) = \eta^*$. The frozen dynamics preserve every policy mass, so $\zeta(\mathbb{R}) \subseteq \mathcal{E}_{\nu^*}$. Corollary 31 therefore gives

$$\eta^* = \zeta(0) = \bar{\eta}^{\nu^*} \in \mathcal{E}^*_{\mathcal{S}\times\Pi}.$$

To obtain the uniform conclusion, suppose it fails. Then there are $\delta > 0$, revision rates $\kappa_n \downarrow 0$, and omega-limit points η_n^∞ of solutions as above with $\mathrm{dist}(\eta_n^\infty, \mathcal{E}^*_{\mathcal{S}\times\Pi}) \ge \delta$. Compactness gives a convergent subsequence. Its limit belongs to $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ by the preceding argument, contradicting continuity of the distance. This proves the limit uniformly over all solutions and their omega-limit points. Theorem 9(i) identifies the aggregate state–action population of each cluster point as a stationary MFG equilibrium population.

On $\mathcal{E}^*_{\mathcal{S}\times\Pi}$, Theorem 5(ii) gives $\lambda(\mu) = (1/\tau_i^*)_i$ and $r \cdot x(\mu)/m = g^*$. Uniform continuity on the compact state–policy simplex therefore yields

$$\lim_{\kappa\downarrow0} \sup_{\eta^\kappa(\cdot)} \sup_{\eta^\infty\in\omega(\eta^\kappa)} \left[\|\lambda(\mu(\eta^\infty)) - (1/\tau_i^*)_i\| + \left| \frac{r \cdot x(\mu(\eta^\infty))}{m} - g^* \right| \right] = 0. \tag{63}$$

□

D.5 State–policy dynamics with exogenous BNN and replicator revision

We replace the common best-response redraw in (14) by Brown–von Neumann–Nash (BNN) or replicator revision, using the same stationary payoffs $H(\nu)$ evaluated at $\bar{\mu}^\nu$. The following observation will also apply to the event-driven rules in Appendix E.6.

Lemma 32 (Convergence when revision vanishes on limiting policy fibers)**.** *Under Assumptions 1 and 4, suppose the autonomous equation $\dot{\eta} = F_{\mathrm{phys}}(\eta) + F_{\mathrm{rev}}(\eta)$ is locally Lipschitz and preserves $\mathcal{M}_m(\mathcal{S} \times \Pi)$, where F_{phys} is the frozen field in* (51) *and F_{rev} is the revision field with its fixed revision coefficient included. Let $Z \subseteq \mathcal{M}_m(\Pi)$ be nonempty and closed, and suppose $F_{\mathrm{rev}} = 0$ on $\mathcal{E}_\nu$ for every $\nu \in Z$. Every solution with $\mathrm{dist}(\nu(t), Z) \to 0$ satisfies*

$$\mathrm{dist}\big(\eta(t), \{\bar{\eta}^\nu : \nu \in Z\}\big) \longrightarrow 0.$$

In particular, if $\nu(t) \to \nu^\infty$ and $F_{\mathrm{rev}} = 0$ on $\mathcal{E}_{\nu^\infty}$, then $\eta(t) \to \bar{\eta}^{\nu^\infty}$.

Proof. Fix $\eta^\infty \in \omega(\eta)$, choose $t_n \ge n$ with $\eta(t_n) \to \eta^\infty$, and put $\eta_n(t) = \eta(t_n + t)$ for $|t| \le n$. The autonomous field is bounded on the compact state–policy simplex, so these shifted solutions have a common derivative bound. Lemma 17(i)–(ii) gives, after relabeling a subsequence, local uniform convergence on $\mathbb{R}$ to a complete solution ζ, with $\zeta(0) = \eta^\infty$. For every fixed $t \in \mathbb{R}$, continuity of the marginal map gives

$$\nu(t_n + t) \longrightarrow \left(\sum_s \zeta_{s,\pi}(t) \right)_{\pi\in\Pi}.$$

Since $t_n + t \to \infty$, the assumed distance limit and closedness of Z place this limiting marginal in Z. Thus $F_{\mathrm{rev}}(\zeta(t)) = 0$, and ζ solves the frozen state–policy dynamics with one constant marginal $\nu^\infty := \nu(\eta^\infty) \in Z$. Corollary 31 implies $\zeta \equiv \bar{\eta}^{\nu^\infty}$. The map $\nu \mapsto \bar{\eta}^\nu$ is continuous by (16) and Lemma 23. Since Z is a closed subset of the compact policy simplex, the target graph $\{\bar{\eta}^\nu : \nu \in Z\}$ is compact. We have placed every point of $\omega(\eta)$ in this graph, so compactness of the state–policy simplex proves the distance limit. Taking $Z = \{\nu^\infty\}$ proves the last assertion. □

Revision rules. Write $\bar{H}(\nu) = m^{-1} \sum_\pi \nu_\pi H_\pi(\nu)$, $b_\pi(\nu) = [H_\pi(\nu) - \bar{H}(\nu)]^+$, and $b = (b_\pi)_{\pi \in \Pi}$. Use the global upper payoff bound $c_{H,\max}$ from the proof of Corollary 26, and choose constants

$$0 < c_{\mathrm{BNN}} \le \frac{1}{|\Pi| c_{H,\max}}, \qquad 0 < c_{\mathrm{rep,exo}} \le \frac{1}{c_{H,\max}}. \tag{64}$$

At a Poisson revision opportunity of rate $\kappa > 0$, BNN revision chooses policy π' with probability $c_{\mathrm{BNN}} b_{\pi'}(\nu)$, and otherwise keeps the incumbent policy. Choosing the incumbent also leaves the policy unchanged. We implement replicator policy dynamics through pairwise proportional imitation: an agent samples π' with probability $\nu_{\pi'}/m$ and adopts it with probability $c_{\mathrm{rep,exo}}[H_{\pi'}(\nu) - H_\pi(\nu)]^+$, where π is the incumbent; otherwise the agent keeps π. Both rules preserve the physical state and immediately use the selected policy's action there. The bounds in (64) make these probabilities valid. The respective state–policy equations are

$$\dot{\eta}_{s,\pi} = (F_{\mathrm{phys}})_{s,\pi}(\eta) + \kappa c_{\mathrm{BNN}} \left[m_s(\eta) b_\pi - \eta_{s,\pi} \sum_{\pi'} b_{\pi'} \right], \tag{65}$$

$$\dot{\eta}_{s,\pi} = (F_{\mathrm{phys}})_{s,\pi}(\eta) + \frac{\kappa c_{\mathrm{rep,exo}}}{m} \sum_{\pi'} \Big(\eta_{s,\pi'} \nu_\pi [H_\pi - H_{\pi'}]^+ - \eta_{s,\pi} \nu_{\pi'} [H_{\pi'} - H_\pi]^+ \Big). \tag{66}$$

All payoffs and excesses here are evaluated at ν. Summing over states gives

$$\dot{\nu}_\pi = \kappa c_{\mathrm{BNN}} m \left[b_\pi - \frac{\nu_\pi}{m} \sum_{\pi'} b_{\pi'} \right], \tag{67}$$

$$\dot{\nu}_\pi = \kappa c_{\mathrm{rep,exo}} \nu_\pi (H_\pi - \bar{H}). \tag{68}$$

In slow time $\widetilde{t} = \kappa t$, these are the BNN and replicator fields of Corollary 26, multiplied by $c_{\mathrm{BNN}} m$ and $c_{\mathrm{rep,exo}}$, respectively.

Proposition 33 (Convergence under exogenous BNN and replicator revision)**.** *Under Assumptions 1 and 4, fix the constants in* (64) *and any $\kappa > 0$. Each of* (65) *and* (66) *has a unique global continuously differentiable solution from every initial point of $\mathcal{M}_m(\mathcal{S} \times \Pi)$.*

(i) BNN revision. *Every solution of* (65) *satisfies*

$$\mathrm{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0, \qquad \mathrm{dist}(\eta(t), \mathcal{E}^*_{\mathcal{S}\times\Pi}) \to 0. \tag{69}$$

The stationary state–policy populations of (65) *are exactly $\mathcal{E}^*_{\mathcal{S}\times\Pi}$.*

(ii) Replicator revision. *Every solution of* (66) *satisfies*

$$\nu(t) \to \nu^\infty \in \arg\max_{\nu' \in \mathcal{E}_0(\nu)} \Phi(\nu'), \qquad \eta(t) \to \bar{\eta}^{\nu^\infty}, \tag{70}$$

where $\mathcal{E}_0(\nu)$ is the initial invariant face in Corollary 26. If $\mathcal{E}_0(\nu) \cap \mathcal{E}^*_\Pi \neq \varnothing$, then $\nu^\infty \in \mathcal{E}^*_\Pi$ and $\bar{\eta}^{\nu^\infty} \in \mathcal{E}^*_{\mathcal{S}\times\Pi}$. Full initial policy support is sufficient for this condition. The stationary state–policy populations are exactly $\bar{\eta}^\nu$ with equal payoffs on $\mathrm{supp}\,\nu$*. Moreover, $\mu(\eta(t)) \to \bar{\mu}^{\nu^\infty}$.*

For BNN revision in (i), and for replicator revision in (ii) when $\mathcal{E}_0(\nu) \cap \mathcal{E}^*_\Pi \neq \varnothing$, the MFG conclusions are

$$\mathrm{dist}(\mu(\eta(t)), \mathcal{E}^*_{\mathcal{I}}) \to 0, \qquad \lambda(\mu(\eta(t))) \to (1/\tau^*_i)_{i \in \mathcal{I}}, \qquad \frac{r \cdot x(\mu(\eta(t)))}{m} \to g^*. \tag{71}$$

Proof. Proposition 2 and Theorem 8(i) give local Lipschitz continuity of both fields. Each field conserves total mass. At $\eta_{s,\pi} = 0$, all outflow terms vanish and all remaining terms are nonnegative. The state–policy simplex is invariant, and compactness gives global existence and uniqueness.

(i) BNN revision. Summing (65) over states gives (67). Corollary 26(iii) therefore gives $\operatorname{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$. For every $\nu^* \in \mathcal{E}^*_\Pi$, the population-weighted average equals the maximum payoff, so $b = 0$ throughout $\mathcal{E}_{\nu^*}$. Lemma 32, with $Z = \mathcal{E}^*_\Pi$, proves (69).

A stationary population must lie in $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ by this convergence. Conversely, every $\bar\eta^{\nu^*}$ with $\nu^* \in \mathcal{E}^*_\Pi$ has zero frozen drift and zero BNN revision. Theorem 9 identifies the aggregate populations in $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ with $\mathcal{E}^*_{\mathcal{I}}$ and gives their common completion rates and reward. Uniform continuity on the compact state–policy simplex therefore gives (71).

(ii) Replicator revision. Summing (66) over states and using $[a]^+ - [-a]^+ = a$ gives (68). Corollary 26(ii) gives $\nu(t) \to \nu^\infty$, where ν^∞ maximizes Φ on $\mathcal{E}_0(\nu)$. If $\nu^\infty_\pi, \nu^\infty_{\pi'} > 0$, transferring sufficiently small mass between these coordinates in either direction stays in that face. The first-order conditions in the two directions therefore give

$$0 = \frac{\partial\Phi}{\partial\nu_\pi}(\nu^\infty) - \frac{\partial\Phi}{\partial\nu_{\pi'}}(\nu^\infty) = \log\frac{H_\pi(\nu^\infty)}{H_{\pi'}(\nu^\infty)}.$$

Thus the payoffs coincide on $\operatorname{supp}\nu^\infty$. For every $\eta \in \mathcal{E}_{\nu^\infty}$,

$$\eta_{s,\pi}\nu^\infty_{\pi'}[H_{\pi'}(\nu^\infty) - H_\pi(\nu^\infty)]^+ = 0. \tag{72}$$

Indeed, if $\nu^\infty_\pi = 0$, then $\sum_s \eta_{s,\pi} = \nu^\infty_\pi = 0$, so nonnegativity gives $\eta_{s,\pi} = 0$ for every s. If $\nu^\infty_{\pi'} = 0$, the factor $\nu^\infty_{\pi'}$ itself is zero. If both $\nu^\infty_\pi > 0$ and $\nu^\infty_{\pi'} > 0$, then $H_{\pi'}(\nu^\infty) = H_\pi(\nu^\infty)$, so the positive-part factor is zero. Lemma 32, with $Z = \{\nu^\infty\}$, proves (70). If the initial face intersects $\mathcal{E}^*_\Pi$, its maximum potential equals the global maximum, so its maximizers lie in $\mathcal{E}^*_\Pi$.

By (68), a policy marginal ν is stationary if and only if the payoffs coincide on $\operatorname{supp}\nu$. For such a marginal, the argument proving (72) makes every revision flux vanish on $\mathcal{E}_\nu$. Corollary 31 then shows that a stationary state–policy population with this marginal must equal $\bar\eta^\nu$. Conversely, every such population has zero frozen drift and zero revision, and is stationary. Continuity of aggregation gives $\mu(\eta(t)) \to \bar\mu^{\nu^\infty}$. Under the initial-face condition, $\nu^\infty \in \mathcal{E}^*_\Pi$, so Theorem 9 and continuity of completion rates and rewards give (71). □

The support condition for replicator revision cannot be dropped. With one state s, two actions a_1, a_2, $m = 1$, $P^{a_j}_{ss} = 1$, $G(x) = x_{sa_1} + x_{sa_2}$, and rewards $r_{sa_1} = 2$, $r_{sa_2} = 1$, the policies $\pi_j(s) = a_j$ have constant payoffs 2 and 1. Starting from $(\nu_{\pi_1}, \nu_{\pi_2}) = (0, 1)$, replicator dynamics stay at the inferior pure policy. At every fixed positive revision rate, BNN trajectories approach the equilibrium set, while replicator trajectories converge to a stationary state–policy population associated with a maximizer of Φ on the initial invariant policy face.

E Proofs and Technical Details for Section 6

We establish existence of solutions for the event-driven state–policy dynamics and prove convergence of the policy marginal. We then identify the cluster points of omega-limit populations as revisions become rare, and prove convergence for any fixed positive revision probability when there exists a unique strict pure policy equilibrium. Appendix E.6 treats completion-weighted BNN revision and replicator revision with a source-rate correction.

E.1 Existence and marginal equations

For the appendix proofs, abbreviate the state–policy and state arrival rates before policy revision by

$$A_{s,\pi}(\eta) := \sum_{s'} X_{s',\pi}(\eta) P^{\pi(s')}_{s's}, \qquad A_s(\eta) := \sum_{\pi'} A_{s,\pi'}(\eta) = \sum_{s',a} x_{s'a}(\mu) P^a_{s's}.$$

Proposition 34. *Under Assumptions 1 and 4, for every $\varepsilon \in [0,1]$, the event-driven state–policy dynamics* (17) *have the following properties.*

(i) For every initial condition in $\mathcal{M}_m(\mathcal{S}\times\Pi)$, the differential inclusion admits at least one global absolutely continuous solution. If a single-valued locally Lipschitz selection $\gamma(\nu) \in \Gamma(\nu)$ is fixed, then the solution is unique.

(ii) Total mass is conserved and nonnegativity is preserved: for all $t \geq 0$,

$$\sum_{s,\pi} \eta_{s,\pi}(t) = m, \qquad \eta_{s,\pi}(t) \geq 0 \quad \textit{for all } (s,\pi) \in \mathcal{S} \times \Pi.$$

(iii) For some measurable selection $\gamma(t) \in \Gamma(\nu(t))$, the policy marginal satisfies, for almost every t,

$$\dot{\nu}_\pi(t) = \varepsilon \left(\frac{\gamma_\pi(t)}{m} \Lambda(\eta(t)) - \chi_\pi(\eta(t)) \right), \qquad \gamma(t) \in \Gamma(\nu(t)). \tag{73}$$

(iv) With the same measurable selection $\gamma(t)$ as in (iii), the state–action population satisfies, for almost every t,

$$\dot{\mu}_{sa} = (1-\varepsilon) \sum_{\pi:\pi(s)=a} A_{s,\pi}(\eta) + \varepsilon \frac{A_s(\eta)}{m} \sum_{\pi:\pi(s)=a} \gamma_\pi - x_{sa}(\mu). \tag{74}$$

This formula also applies when $A_s(\eta) = 0$.

Proof. For $\gamma \in \Gamma(\nu)$, an Euler step with $0 < \Delta t \leq \overline{\lambda}^{-1}$ is

$$\eta^+_{s,\pi} = [1 - \Delta t \lambda_{s,\pi(s)}(\mu)]\eta_{s,\pi} + \Delta t(1-\varepsilon)A_{s,\pi} + \Delta t \varepsilon \frac{\gamma_\pi}{m} A_s \geq 0.$$

Since $\sum_s A_{s,\pi} = \chi_\pi$ and $\sum_\pi \chi_\pi = \sum_s A_s = \Lambda$,

$$\sum_{s,\pi} \eta^+_{s,\pi} = m + \Delta t[(1-\varepsilon)\Lambda + \varepsilon\Lambda - \Lambda] = m.$$

Lemma 24 gives global existence. If a single-valued locally Lipschitz selection $\gamma(\nu) \in \Gamma(\nu)$ is fixed, the right-hand side of (17) is locally Lipschitz. Uniqueness then follows from the standard ODE uniqueness theorem (Teschl, 2012, Theorem 2.2).

Summing over states gives the policy marginal equation

$$\dot{\nu}_\pi = (1-\varepsilon)\chi_\pi + \varepsilon \frac{\gamma_\pi}{m}\Lambda - \chi_\pi = \varepsilon \left(\frac{\gamma_\pi}{m}\Lambda - \chi_\pi \right).$$

Summing over policies then gives $\sum_\pi \dot{\nu}_\pi = 0$. Since $\Lambda \geq m\underline{\lambda} > 0$, for $\varepsilon > 0$ the selection can be recovered as

$$\gamma_\pi = \frac{m}{\Lambda} \left(\frac{\dot{\nu}_\pi}{\varepsilon} + \chi_\pi \right),$$

and is measurable. For $\varepsilon = 0$, choose any measurable selection $\gamma(t) \in \Gamma(\nu(t))$. At a boundary point with $\eta_{s,\pi} = 0$, (17) gives $\dot{\eta}_{s,\pi} = (1-\varepsilon)A_{s,\pi} + \varepsilon(\gamma_\pi/m)A_s \geq 0$. Thus the nonnegative orthant is forward invariant. With the same measurable selection $\gamma(t)$,

$$\dot{\mu}_{sa} = \sum_{\pi:\pi(s)=a} \dot{\eta}_{s,\pi} = (1-\varepsilon) \sum_{\pi:\pi(s)=a} A_{s,\pi} + \varepsilon \frac{A_s}{m} \sum_{\pi:\pi(s)=a} \gamma_\pi - \underbrace{\sum_{\pi:\pi(s)=a} X_{s,\pi}}_{x_{sa}(\mu)}.$$

This is (74), including $A_s = 0$. Finally, $\|\dot{\eta}\|_1 \leq 2\Lambda \leq 2m\overline{\lambda}$ permits continuation at every finite endpoint. □

Policy balance and stationarity of the state–policy population. Fix $\nu^* \in \mathcal{E}^*_\Pi$ and $\varepsilon \in (0,1]$. Recall from Section 6 that $e(\eta) = (\chi_\pi(\eta)/\Lambda(\eta))_{\pi\in\Pi}$ is the vector of policy shares of completion events. At $\eta = \bar{\eta}^{\nu^*}$, the state–policy population is stationary under the frozen state–policy dynamics (Proposition 10): the arrival rate satisfies $A_{s,\pi}(\bar{\eta}^{\nu^*}) = X_{s,\pi}(\bar{\eta}^{\nu^*}) = \bar{X}^{\nu^*}_{s,\pi}$. Positive completion rates give $e_\pi(\bar{\eta}^{\nu^*}) > 0$ exactly when $\nu^*_\pi > 0$. Since $\nu^* \in \mathcal{E}^*_\Pi$, the vector $me(\bar{\eta}^{\nu^*})$ has mass m and is supported on best responses, so it lies in $\Gamma(\nu^*)$. Equation (18) forces this choice at stationarity:

$$\dot{\nu} = 0 \quad \Longleftrightarrow \quad \gamma_\pi = m\frac{\chi_\pi(\bar{\eta}^{\nu^*})}{\Lambda(\bar{\eta}^{\nu^*})} = me_\pi(\bar{\eta}^{\nu^*}) \quad \forall \pi.$$

Substitution in (17) now gives

$$\dot{\eta}_{s,\pi} = \varepsilon\big[e_\pi(\bar{\eta}^{\nu^*})A_s(\bar{\eta}^{\nu^*}) - A_{s,\pi}(\bar{\eta}^{\nu^*})\big].$$

Since $\varepsilon > 0$, stationarity is therefore equivalent to

$$\dot{\eta} = 0 \quad \Longleftrightarrow \quad A_{s,\pi}(\bar{\eta}^{\nu^*}) = e_\pi(\bar{\eta}^{\nu^*})A_s(\bar{\eta}^{\nu^*}) \quad \forall (s,\pi).$$

Since $A_{s,\pi}(\bar{\eta}^{\nu^*}) = \bar{X}^{\nu^*}_{s,\pi}$ and $A_s(\bar{\eta}^{\nu^*}) = \sum_\sigma \bar{X}^{\nu^*}_{s,\sigma}$, this is equivalent to

$$\bar{X}^{\nu^*}_{s,\pi} = e_\pi(\bar{\eta}^{\nu^*}) \sum_\sigma \bar{X}^{\nu^*}_{s,\sigma}, \qquad (s,\pi) \in \mathcal{S} \times \Pi. \tag{75}$$

The right-hand side is the total completion rate at state s multiplied by the probability of selecting policy π; it must match that policy's completion rate at s. For $A_s(\bar{\eta}^{\nu^*}) > 0$, division by $A_s(\bar{\eta}^{\nu^*})$ gives

$$\frac{A_{s,\pi}(\bar{\eta}^{\nu^*})}{A_s(\bar{\eta}^{\nu^*})} = \frac{\chi_\pi(\bar{\eta}^{\nu^*})}{\Lambda(\bar{\eta}^{\nu^*})}.$$

Thus the policy share of arrivals at state s equals its share of all completion events. If $A_s(\bar{\eta}^{\nu^*}) = 0$, nonnegativity gives $A_{s,\pi}(\bar{\eta}^{\nu^*}) = 0$ for every π, so the condition is automatic.

E.2 Frozen state–policy dynamics in event time

We first justify the use of event time and transfer the convergence results for the frozen state–policy dynamics to this time variable. We then bound the revision term and trajectory derivatives for the compactness argument.

Recall the event clock introduced in Section 6:

$$\theta(t) := \int_0^t \Lambda(\eta(t'))\, dt'. \tag{76}$$

Since $\Lambda(\eta) > 0$, it is strictly increasing. Dividing (17) by $\Lambda(\eta)$ gives the event-driven state–policy dynamics in event time for $\hat{\eta}(\theta) := \eta(t(\theta))$, $\hat{\nu}(\theta) := \nu(t(\theta))$, and $\hat{\gamma}(\theta) := \gamma(t(\theta))$:

$$\frac{d\hat{\eta}_{s,\pi}}{d\theta} = \frac{(1-\varepsilon)A_{s,\pi}(\hat{\eta}) - X_{s,\pi}(\hat{\eta})}{\Lambda(\hat{\eta})} + \varepsilon \frac{\hat{\gamma}_\pi}{m} \frac{A_s(\hat{\eta})}{\Lambda(\hat{\eta})}. \tag{77}$$

Here $\hat{\gamma}(\theta) \in \Gamma(\hat{\nu}(\theta))$, and the policy marginal is (19). At $\varepsilon = 0$,

$$\frac{d\hat{\eta}_{s,\pi}}{d\theta} = \frac{A_{s,\pi}(\hat{\eta}) - X_{s,\pi}(\hat{\eta})}{\Lambda(\hat{\eta})}, \tag{78}$$

which is the frozen state–policy dynamics (51) expressed in event time. At $\eta = \bar{\eta}^\nu$, the completion shares are

$$e_\pi(\bar{\eta}^\nu) = \frac{\bar{\chi}^\nu_\pi}{\sum_{\sigma\in\Pi} \bar{\chi}^\nu_\sigma}. \tag{79}$$

We first verify that the change from physical time to event time preserves limit sets and the sign of derivatives of Lyapunov functions.

Lemma 35. *Under Assumptions 1 and 4, for every state–policy population $\eta \in \mathcal{M}_m(\mathcal{S} \times \Pi)$,*

$$0 < m\underline{\lambda} \le \Lambda(\eta) \le m\overline{\lambda} < \infty. \tag{80}$$

Consequently the clock θ in (76) *and its inverse are Lipschitz. They map a forward trajectory on $[0,\infty)$ to a forward trajectory on $[0,\infty)$, preserving its omega-limit set. For a complete trajectory, both clocks range over $\mathbb{R}$, and the alpha-limit set is also preserved. For a Lyapunov function L, the derivatives dL/dt and $dL/d\theta$ have the same sign and vanish simultaneously at corresponding times, wherever they exist.*

Proof. The rate bounds and total mass give

$$m\underline{\lambda} \le \Lambda(\eta) = \sum_{s,\pi} \eta_{s,\pi}\lambda_{s,\pi(s)}(\mu) \le m\overline{\lambda}.$$

Integrating, for $t_1 < t_2$,

$$m\underline{\lambda}(t_2 - t_1) \le \theta(t_2) - \theta(t_1) \le m\overline{\lambda}(t_2 - t_1).$$

Thus the clock and its inverse are Lipschitz, and $t \to \infty$ if and only if $\theta \to \infty$. For a complete trajectory, the same bounds give $t \to -\infty$ if and only if $\theta \to -\infty$. The reparameterized trajectory visits the same population vectors in the same order, so its omega-limit set, and its alpha-limit set when defined, agree with those of the original trajectory. The chain rule gives

$$\frac{d}{d\theta}L(\hat{\eta}(\theta)) = \frac{1}{\Lambda(\hat{\eta}(\theta))}\left.\frac{d}{dt}L(\eta(t))\right|_{t=t(\theta)}.$$

Since $\Lambda(\hat{\eta}(\theta)) > 0$, the two derivatives have the same sign, and one is zero if and only if the other is zero. □

We omit hats when the use of event time θ is clear. The change to event time also transfers Corollary 31: every complete solution of the frozen dynamics is stationary.

Corollary 36. *Under Assumptions 1 and 4, every complete solution η of* (78)*, defined for all $\theta \in \mathbb{R}$, with policy marginal ν is constant and equals $\bar{\eta}^{\nu}$. In particular it has zero transient state–policy mass, and its population masses satisfy $\eta_{s,\pi} > 0$ whenever $\nu_\pi > 0$ and $s \in \mathcal{S}^{\pi}_{\mathrm{rec}}$.*

Proof. Lemma 35 converts the trajectory into a complete solution of the frozen state–policy dynamics in physical time on $\mathcal{E}_\nu$. Corollary 31 identifies it as $\bar{\eta}^{\nu}$, where

$$\bar{\eta}^{\nu}_{s,\pi} = \bar{\chi}^{\nu}_{\pi} q^{\pi}_{s} \partial_{s,\pi(s)} G(\bar{x}^{\nu}).$$

Since $\partial_i G > 0$ and $q^{\pi}_{s} > 0$ exactly on $\mathcal{S}^{\pi}_{\mathrm{rec}}$,

$$\bar{\eta}^{\nu}_{s,\pi} > 0 \quad \Longleftrightarrow \quad \nu_\pi > 0 \text{ and } s \in \mathcal{S}^{\pi}_{\mathrm{rec}}.$$

□

We now restore policy revision. The following decomposition separates the frozen field from the revision term and gives the bounds used in the compactness argument. For a measurable best-response selection $\gamma(\theta) \in \Gamma(\nu(\theta))$, write (77) as

$$\frac{d\eta}{d\theta} = F_{\mathrm{evt}}(\eta) + \varepsilon F^{\mathrm{evt}}_{\mathrm{rev}}(\eta, \gamma), \tag{81}$$

where

$$(F_{\mathrm{evt}})_{s,\pi}(\eta) := \frac{A_{s,\pi}(\eta) - X_{s,\pi}(\eta)}{\Lambda(\eta)}, \tag{82}$$

$$(F^{\mathrm{evt}}_{\mathrm{rev}})_{s,\pi}(\eta, \gamma) := \frac{\gamma_\pi A_s(\eta)/m - A_{s,\pi}(\eta)}{\Lambda(\eta)}. \tag{83}$$

Thus $F_{\mathrm{evt}} = F_{\mathrm{phys}}/\Lambda$, with F_{phys} defined by (51); these are the vector fields of the frozen state–policy dynamics in event time and physical time, respectively. For this revision rule, the physical-time revision field is $F_{\mathrm{rev}}(\eta) = \varepsilon\Lambda(\eta)F^{\mathrm{evt}}_{\mathrm{rev}}(\eta, \gamma)$ when the selection γ is fixed as a function of the population. The field F_{evt} is globally Lipschitz on the compact state–policy simplex and $F^{\mathrm{evt}}_{\mathrm{rev}}$ is uniformly bounded, by Proposition 2 and Lemma 35. More explicitly, $\sum_{s,\pi} A_{s,\pi} = \sum_{s,\pi} X_{s,\pi} = \sum_{s,\pi} \gamma_\pi A_s/m = \Lambda$ gives $\|F_{\mathrm{evt}}(\eta)\|_1 \le 2$ and $\|F^{\mathrm{evt}}_{\mathrm{rev}}(\eta, \gamma)\|_1 \le 2$. For the event-driven state–policy dynamics (77), nonnegative inflow and outflow rates give

$$\left\|\frac{d\eta}{d\theta}\right\|_1 \le \frac{\sum_{s,\pi}[(1-\varepsilon)A_{s,\pi} + \varepsilon\gamma_\pi A_s/m + X_{s,\pi}]}{\Lambda} = 2 \quad (0 \le \varepsilon \le 1). \tag{84}$$

E.3 Proof of Proposition 12: event-driven policy convergence

The policy potential (10) is evaluated along (17); no stationarity assumption on the state–policy population is used in its derivative. The uniform rate bounds from Section 6 turn the derivative identity (20) into convergence of the policy marginal.

Proof of Proposition 12. Theorem 8(i) and (19) give

$$\begin{aligned}\frac{d}{d\theta}\Phi(\nu) &= \varepsilon \sum_{\pi} (\log H_\pi - 1)\left(\frac{\gamma_\pi}{m} - e_\pi\right) \\ &= \varepsilon \left(\max_{\sigma} \log H_\sigma - \sum_{\pi} e_\pi \log H_\pi\right) \\ &= \varepsilon \sum_{\pi} e_\pi \log \frac{\max_\sigma H_\sigma}{H_\pi} \geq 0.\end{aligned}$$

Here $\sum_\pi \gamma_\pi/m = \sum_\pi e_\pi = 1$, and γ is supported on best responses. The rate bounds imply

$$\underline{\lambda}\nu_\pi \leq \chi_\pi \leq \overline{\lambda}\nu_\pi.$$

Dividing by Λ and using (80) gives

$$\frac{\underline{\lambda}}{m\overline{\lambda}}\nu_\pi \leq e_\pi \leq \frac{\overline{\lambda}}{m\underline{\lambda}}\nu_\pi.$$

Thus $\operatorname{supp} e = \operatorname{supp} \nu$, and the bracket in (20) vanishes exactly on $\mathcal{E}^*_\Pi$. For $\varepsilon > 0$, it follows that $d\Phi/d\theta = 0$ exactly when $\nu \in \mathcal{E}^*_\Pi$. Returning to physical time,

$$\frac{d}{dt}\Phi(\nu(t)) = \varepsilon \sum_{\pi} \chi_\pi \log \frac{\max_\sigma H_\sigma}{H_\pi} \geq \varepsilon \underline{\lambda} M_{\mathrm{BR}}(\nu) \geq \varepsilon \underline{\lambda} \Delta(\nu).$$

Hence $\Delta(\nu(t)) \leq e^{-\varepsilon \underline{\lambda} t}\Delta(\nu(0))$. The lower bound for e_π similarly gives

$$\Delta(\nu(\theta)) \leq \exp\left(-\frac{\varepsilon\underline{\lambda}}{m\overline{\lambda}}\theta\right)\Delta(\nu(0)). \tag{85}$$

On the compact policy simplex, $\Delta \geq 0$ is continuous and $\Delta^{-1}(\{0\}) = \mathcal{E}^*_\Pi$; therefore

$$\min_{\operatorname{dist}(\nu, \mathcal{E}^*_\Pi) \geq \delta} \Delta(\nu) > 0 \quad (\delta > 0 \text{ whenever this set is nonempty}).$$

This positive lower bound and (85) prove convergence to $\mathcal{E}^*_\Pi$ and Lyapunov stability, uniformly over state–policy populations with the given policy marginal. □

E.4 Proof of Theorem 13: rare event-driven revisions

Proposition 12 gives convergence of the policy marginal to $\mathcal{E}^*_\Pi$. We then consider omega-limit points of the state–policy trajectories as $\varepsilon \downarrow 0$. Trajectory compactness shows that every cluster point lies on a complete solution of the frozen state–policy dynamics. Since every such solution is stationary, each cluster point equals $\bar{\eta}^{\nu^*}$ for some $\nu^* \in \mathcal{E}^*_\Pi$.

Proof of Theorem 13. Global existence is Proposition 34. Lemma 35 gives $\theta(t) \to \infty$ as $t \to \infty$ and shows that the time change preserves omega-limit sets. Proposition 12 therefore places the policy marginal of every point of $\omega(\hat{\eta}^\varepsilon)$ in $\mathcal{E}^*_\Pi$.

Let $\varepsilon_n \downarrow 0$, let $\hat{\eta}^{\varepsilon_n}$ be arbitrary solutions in event time from initial populations in $\mathcal{M}_m(\mathcal{S} \times \Pi)$, and choose $\hat{\eta}^\infty_n \in \omega(\hat{\eta}^{\varepsilon_n})$. Consider any convergent subsequence $\hat{\eta}^\infty_n \to \eta^*$, relabeled by n. Continuity of the marginal map and closedness of $\mathcal{E}^*_\Pi$ imply that the marginal ν^* of η^* lies in $\mathcal{E}^*_\Pi$. Choose $\theta_n \geq n$ with $\|\hat{\eta}^{\varepsilon_n}(\theta_n) - \hat{\eta}^\infty_n\| \leq n^{-1}$, and set

$$\zeta_n(\theta) := \hat{\eta}^{\varepsilon_n}(\theta_n + \theta), \qquad |\theta| \leq n.$$

The shifted trajectories lie in the compact state–policy simplex, $\zeta_n(0) \to \eta^*$, and $\|d\zeta_n/d\theta\|_1 \le 2$ by (84). The decomposition (81) and the bound $\|F^{\mathrm{evt}}_{\mathrm{rev}}(\eta,\gamma)\|_1 \le 2$ give, for each fixed $T < \infty$ and $n \ge T$,

$$\sup_{a,b\in[-T,T]} \left\| \zeta_n(b) - \zeta_n(a) - \int_a^b F_{\mathrm{evt}}(\zeta_n(\theta))\, d\theta \right\| \le 4T\varepsilon_n \longrightarrow 0.$$

Similar to the proof of Theorem 11(ii), we apply Lemma 17(i)–(ii) with $I = \mathbb{R}$ and $f = F_{\mathrm{evt}}$. This gives a further subsequence converging uniformly on every bounded event-time interval to a complete solution ζ of (78), with $\zeta(0) = \eta^*$. The frozen dynamics preserve policy masses, so $\zeta(\mathbb{R}) \subseteq \mathcal{E}_{\nu^*}$. Corollary 36 then gives

$$\eta^* = \zeta(0) = \bar{\eta}^{\nu^*} \in \mathcal{E}^*_{\mathcal{S}\times\Pi}.$$

If the asserted uniform limit failed, there would be a sequence of such omega-limit points whose distances from $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ are bounded below by some $\delta > 0$. Compactness would give a convergent subsequence, whose limit belongs to $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ by the preceding argument, contradicting continuity of the distance. Theorem 9(i) identifies the corresponding stationary MFG equilibrium populations.

Finally, the completion-rate and reward maps are uniformly continuous on the compact state–policy simplex and have the common values $(1/\tau_i^*)_i$ and g^* on $\mathcal{E}^*_{\mathcal{S}\times\Pi}$, so

$$\lim_{\varepsilon\downarrow 0} \sup_{\hat{\eta}^\varepsilon(\cdot)} \sup_{\hat{\eta}^\infty \in \omega(\hat{\eta}^\varepsilon)} \left[\|\lambda(\mu(\hat{\eta}^\infty)) - (1/\tau_i^*)_i\| + \left| \frac{r \cdot x(\mu(\hat{\eta}^\infty))}{m} - g^* \right| \right] = 0.$$

□

Remark 37. Convergence of the state–action population to $\mathcal{E}^*_{\mathcal{I}}$ does not hold in general at a fixed positive revision probability. To see this, take $m = 1$, $\mathcal{S} = \mathcal{A} = \{0,1\}$, $r_{sa} = 1$, $G(x) = \sum_{s,a} x_{sa}$, and

$$P^0 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad P^1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

The model is weakly communicating, all completion rates and policy payoffs equal one, and $\mathcal{E}^*_{\Pi}$ is the whole policy simplex. The unichain policies are $\pi^{00}, \pi^{11}, \pi^{10}$; the identity policy π^{01} is multichain. The two superscript digits specify the actions at states 0 and 1, respectively. Use only the first two policies, with zero mass and zero probability of selecting π^{10}. Fix $0 < \varepsilon \le 1$ and $0 < c_{\mathrm{amp}} < 1/(2\sqrt{2})$, and define for all real t

$$v(t) = \tfrac{1}{2} + c_{\mathrm{amp}} \sin(\varepsilon t), \qquad p(t) = \tfrac{1}{2} + c_{\mathrm{amp}}[\sin(\varepsilon t) + \cos(\varepsilon t)].$$

The revision rule $\gamma(t) = (p(t), 1-p(t), 0)$ is a measurable best-response selection, and $\dot{v} = \varepsilon(p - v)$. With columns indexed by π^{00}, π^{11}, put

$$\mathsf{A}(t) = \begin{pmatrix} v(1-\varepsilon+\varepsilon p) & \varepsilon v(1-p) \\ \varepsilon(1-v)p & (1-v)(1-\varepsilon p) \end{pmatrix}, \qquad \eta(t) = \int_0^\infty e^{-t'} \mathsf{A}(t-t')\, dt'.$$

All entries of A are nonnegative and sum to one, so η is a periodic state–policy trajectory of mass one. Its first column sums to v: the derivative of the column sum plus the sum itself equals $(1-\varepsilon)v + \varepsilon p = \dot{v} + v$, so uniqueness of the periodic solution gives v. Thus A is exactly the inflow matrix in (17), and $\dot{\eta} = \mathsf{A} - \eta$ verifies that equation. The state mass $m_0(t) = \sum_\pi \eta_{0,\pi}(t)$ satisfies

$$m_0(t) = \tfrac{1}{2} + \frac{c_{\mathrm{amp}}}{1+\varepsilon^2}[\sin(\varepsilon t) - \varepsilon\cos(\varepsilon t)], \qquad \dot{m}_0 = v - m_0.$$

Stationary state balance in (3) requires $\mu_{10} = \mu_{01}$, whereas here $\mu_{10} - \mu_{01} = v - m_0$ is nonconstant and periodic. In fact

$$\limsup_{t\to\infty} \operatorname{dist}(\mu(t), \mathcal{E}^*_{\mathcal{I}}) \ge \frac{c_{\mathrm{amp}}\varepsilon}{\sqrt{2(1+\varepsilon^2)}} > 0.$$

The policy marginal remains in $\mathcal{E}^*_{\Pi}$ at all times. Thus policy equilibrium need not imply stationary state–action balance. The state–policy convergence conclusion of Theorem 13 therefore requires $\varepsilon \downarrow 0$.

E.5 State–policy convergence under a strict pure policy equilibrium

A strict pure policy equilibrium is a special case in which the state–policy population converges without sending the revision probability to zero. Write $\mathbf{e}_\pi \in \mathbb{R}^\Pi$ for the standard basis vector indexed by policy π.

Theorem 38. *Under Assumptions 1 and 4, suppose there is a policy π^* with $\mathrm{BR}_\Pi(m\mathbf{e}_{\pi^*}) = \{\pi^*\}$. Then $\mathcal{E}^*_\Pi = \{m\mathbf{e}_{\pi^*}\}$, and for every fixed $\varepsilon \in (0,1]$, every absolutely continuous solution of the event-driven differential inclusion (77) converges to $\bar\eta^{m\mathbf{e}_{\pi^*}}$.*

Proof. Put $\nu^* = m\mathbf{e}_{\pi^*}$. For $\nu \in \mathcal{M}_m(\Pi) \setminus \{\nu^*\}$, strictness gives

$$\Phi(\nu) - \Phi(\nu^*) \le \nabla\Phi(\nu^*)\cdot(\nu-\nu^*) = \sum_{\pi\neq\pi^*} \nu_\pi \log\frac{H_\pi(\nu^*)}{H_{\pi^*}(\nu^*)} < 0.$$

Thus $\mathcal{E}^*_\Pi = \{\nu^*\}$, and Proposition 12 gives $\nu(\theta)\to\nu^*$. Strictness means $H_{\pi^*}(\nu^*) > H_\pi(\nu^*)$ for every $\pi \neq \pi^*$. Since Π is finite and each payoff H_π is continuous, all these inequalities hold throughout some neighborhood of ν^*. Convergence of $\nu(\theta)$ places it in this neighborhood for every $\theta \ge \theta_0$, for some θ_0. Thus $\mathrm{BR}_\Pi(\nu(\theta)) = \{\pi^*\}$, and every best-response selection satisfies $\gamma(\theta) = m\mathbf{e}_{\pi^*}$ almost everywhere after θ_0. By Lemma 35, the corresponding physical time $t_0 = t(\theta_0)$ is finite and $\nu(t)\to\nu^*$. For $t \ge t_0$, equation (17) is the autonomous equation

$$\dot\eta = F_{\mathrm{phys}}(\eta) + F_{\mathrm{rev}}(\eta), \qquad (F_{\mathrm{rev}})_{s,\pi}(\eta) = \varepsilon\left(\frac{\nu^*_\pi}{m}A_s(\eta) - A_{s,\pi}(\eta)\right).$$

This field is locally Lipschitz by Proposition 2. For every state–policy population, its inflows are $(1-\varepsilon)A_{s,\pi} + \varepsilon(\nu^*_\pi/m)A_s \ge 0$, and its outflows are $X_{s,\pi}$. Both sum to Λ, and the outflow vanishes at a zero component. Thus this fixed-probability field preserves the state–policy simplex, including outside the neighborhood where π^* is the unique best response. On $\mathcal{E}_{\nu^*}$, all other policy masses vanish, so $A_{s,\pi} = 0$ for $\pi \neq \pi^*$ and $A_{s,\pi^*} = A_s$. Consequently $F_{\mathrm{rev}} = 0$ on $\mathcal{E}_{\nu^*}$. Lemma 32, applied to the trajectory $t \mapsto \eta(t_0+t)$ with $Z = \{\nu^*\}$, gives $\eta(t)\to\bar\eta^{\nu^*}$. Lemma 35 transfers this limit to event time. □

E.6 Event-driven state–policy dynamics with BNN and replicator revision

We now replace the common best-response redraw in (17) by two rules at completion events. The payoffs remain $H(\nu)$, evaluated at $\bar\mu^\nu$, while completion shares and individual completion rates use the current population η. At a completion from (s,π), the agent first moves to s' with probability $P^{\pi(s)}_{ss'}$, then receives a revision opportunity with probability $\varepsilon \in (0,1]$. All probabilities below use the population immediately before completion. Use the arrival rates $A_{s,\pi}, A_s$ from Appendix E.1, and the payoff bound $c_{H,\max}$ and BNN normalization c_{BNN} from Appendix D.5.

Completion-weighted BNN revision. Define

$$\bar H_{\mathrm{evt}}(\eta) := \sum_\pi e_\pi(\eta)H_\pi(\nu), \qquad b^{\mathrm{evt}}_\pi(\eta) := [H_\pi(\nu) - \bar H_{\mathrm{evt}}(\eta)]^+, \tag{86}$$

and write $b^{\mathrm{evt}} = (b^{\mathrm{evt}}_\pi)_{\pi\in\Pi}$. At an opportunity, choose π' with probability $c_{\mathrm{BNN}}b^{\mathrm{evt}}_{\pi'}$, and otherwise keep the incumbent; choosing the incumbent also leaves the policy unchanged. Since $c_{\mathrm{BNN}}\sum_{\pi'} b^{\mathrm{evt}}_{\pi'} \le 1$, this gives

$$\begin{aligned}\dot\eta_{s,\pi} &= \left(1 - \varepsilon c_{\mathrm{BNN}}\sum_{\pi'} b^{\mathrm{evt}}_{\pi'}\right)A_{s,\pi} - X_{s,\pi} + \varepsilon c_{\mathrm{BNN}} b^{\mathrm{evt}}_\pi A_s,\\ \dot\nu_\pi &= \varepsilon c_{\mathrm{BNN}}\left(\Lambda b^{\mathrm{evt}}_\pi - \chi_\pi\sum_{\pi'} b^{\mathrm{evt}}_{\pi'}\right).\end{aligned} \tag{87}$$

The average payoff $\bar H_{\mathrm{evt}}(\eta)$ weights each policy's payoff by its share $e_\pi(\eta) = \chi_\pi(\eta)/\Lambda(\eta)$ of the current total completion rate. The average payoff $H(\nu)$ instead uses the policy mass shares ν_π/m. These averages agree when $e(\eta) = \nu/m$.

Replicator revision with source-rate correction. We implement replicator policy dynamics at completion events through a source-rate correction to the pairwise proportional imitation rule. Fix $0 < c_{\mathrm{rep,evt}} \le \underline{\lambda}/c_{H,\max}$. At an opportunity following a completion from (s,π), sample π' with probability $\nu_{\pi'}/m$ and adopt it with probability $c_{\mathrm{rep,evt}}[H_{\pi'} - H_\pi]^+/\lambda_{s,\pi(s)}(\mu)$; otherwise keep π. Thus the conditional switching probability is

$$p^{\mathrm{evt}}_{s,\pi\to\pi'}(\eta) := \frac{c_{\mathrm{rep,evt}}}{\lambda_{s,\pi(s)}(\mu)}\frac{\nu_{\pi'}}{m}[H_{\pi'}(\nu) - H_\pi(\nu)]^+. \tag{88}$$

The denominator is the rate of the completed action at its source state, and $\mu = \mu(\eta)$. The bound on $c_{\mathrm{rep,evt}}$ ensures that the adoption probability is at most one and $\sum_{\pi'} p^{\mathrm{evt}}_{s,\pi\to\pi'} \le 1$, even on the boundary. The state–policy equation is

$$\dot\eta_{s,\pi} = \sum_{s'} X_{s',\pi} P^{\pi(s')}_{s's}\left(1 - \varepsilon\sum_{\pi'} p^{\mathrm{evt}}_{s',\pi\to\pi'}\right) - X_{s,\pi} + \varepsilon\sum_{s',\pi'} X_{s',\pi'} P^{\pi'(s')}_{s's} p^{\mathrm{evt}}_{s',\pi'\to\pi}. \tag{89}$$

The source-rate correction gives the exact cancellation

$$\begin{aligned}\varepsilon\sum_{s,s'} X_{s,\pi} P^{\pi(s)}_{ss'} p^{\mathrm{evt}}_{s,\pi\to\pi'} = \varepsilon\sum_s X_{s,\pi} p^{\mathrm{evt}}_{s,\pi\to\pi'} = \frac{\varepsilon c_{\mathrm{rep,evt}}}{m}\nu_\pi\nu_{\pi'}[H_{\pi'} - H_\pi]^+,\\ \dot\nu_\pi = \varepsilon c_{\mathrm{rep,evt}}\nu_\pi(H_\pi - \bar H).\end{aligned} \tag{90}$$

Here $\bar H = m^{-1}\sum_\pi \nu_\pi H_\pi$. The first equality uses row stochasticity to sum over arrival states. For the second, substitute $X_{s,\pi} = \eta_{s,\pi}\lambda_{s,\pi(s)}(\mu)$ in (88) and use $\sum_s \eta_{s,\pi} = \nu_\pi$. Subtracting opposite switching fluxes and using $[a]^+ - [-a]^+ = a$ gives the marginal equation. Thus the policy marginal follows replicator dynamics in physical time, up to a constant change of speed.

For both rules, event time is $\theta(t) = \int_0^t \Lambda(\eta(t'))\,dt'$. Using the same symbols for reparameterized trajectories, the marginals are

$$\frac{d\nu_\pi}{d\theta} = \varepsilon c_{\mathrm{BNN}}\left(b^{\mathrm{evt}}_\pi - e_\pi\sum_{\pi'} b^{\mathrm{evt}}_{\pi'}\right), \tag{91}$$

$$\frac{d\nu_\pi}{d\theta} = \frac{\varepsilon c_{\mathrm{rep,evt}}}{\Lambda(\eta)}\nu_\pi(H_\pi - \bar H), \tag{92}$$

respectively. In slow time $\widetilde\theta = \varepsilon\theta$, the factor ε disappears. The bounds $m\underline{\lambda} \le \Lambda \le m\overline{\lambda}$ ensure that physical and event time tend to infinity together.

Proposition 39 (Convergence under event-driven BNN and replicator revision)**.** *Under Assumptions 1 and 4, fix the normalizing constants above and $\varepsilon \in (0,1]$. Both* (87) *and* (89) *have a unique global continuously differentiable solution from every initial point of $\mathcal{M}_m(\mathcal{S}\times\Pi)$.*

(i) BNN revision. *For completion-weighted BNN,* (69) *holds, and the stationary state–policy populations are exactly $\mathcal{E}^*_{\mathcal{S}\times\Pi}$.*

(ii) Replicator revision. *With the source-rate correction,* (70) *holds, and the stationary state–policy populations are exactly $\bar\eta^\nu$ with equal payoffs on* $\operatorname{supp}\nu$.

The MFG conclusions (71) *hold for BNN revision in (i) from every initial population and for replicator revision in (ii) when $\mathcal{E}_0(\nu)\cap\mathcal{E}^*_\Pi \neq \varnothing$; in the latter case $\mu(\eta(t)) \to \bar\mu^{\nu^\infty} \in \mathcal{E}^*_{\mathcal{I}}$. All limits hold in physical and event time, at the fixed revision probability.*

Proof. Proposition 2 and Theorem 8(i), together with the positive lower bounds on Λ and λ_i, make both fields locally Lipschitz. Each completion produces one arrival distributed among state–policy pairs with nonnegative probabilities summing to one. Thus inflow and outflow both sum to Λ; at a zero component, the outflow is zero and all inflows are nonnegative. The state–policy simplex is invariant, so solutions are unique and global, with $\|\dot\eta(t)\|_1 \le 2\Lambda \le 2m\overline{\lambda}$. Since $t\to\infty$ if and only if $\theta\to\infty$, the convergence conclusions proved below in physical time also hold in event time.

(i) BNN revision. Conservation of total policy mass, $\sum_\pi \nu_\pi = m$, and $\partial\Phi(\nu)/\partial\nu_\pi = \log H_\pi(\nu) - 1$ give

$$\begin{aligned}\frac{d}{d\theta}\Phi(\nu) &= \varepsilon c_{\mathrm{BNN}} \sum_\pi b_\pi^{\mathrm{evt}} \left(\log H_\pi - \sum_{\pi'} e_{\pi'} \log H_{\pi'} \right) \\ &\geq \varepsilon c_{\mathrm{BNN}} \sum_\pi b_\pi^{\mathrm{evt}} \log \frac{H_\pi}{\bar{H}_{\mathrm{evt}}} \geq 0, \end{aligned} \tag{93}$$

by Jensen's inequality. The last sum is positive whenever some $b_\pi^{\mathrm{evt}} > 0$. If every excess is zero, every payoff is at most its completion-weighted average, forcing that average to equal the maximum and every positive-weight policy to attain it. The bounds $\underline{\lambda}\nu_\pi \leq \chi_\pi \leq \overline{\lambda}\nu_\pi$ give $\operatorname{supp} e = \operatorname{supp} \nu$, so

$$\frac{d}{d\theta}\Phi(\nu) = 0 \quad \Longleftrightarrow \quad b^{\mathrm{evt}} = 0 \quad \Longleftrightarrow \quad \nu \in \mathcal{E}_\Pi^*. \tag{94}$$

Conversely, at a policy equilibrium the completion-weighted average uses only maximizing policies, so all excesses vanish. Multiplication by $\Lambda > 0$ gives the same sign and equality condition in physical time.

In physical time, (87) has the form $\dot{\eta} = F_{\mathrm{phys}}(\eta) + F_{\mathrm{rev}}(\eta)$, where

$$(F_{\mathrm{rev}})_{s,\pi}(\eta) = \varepsilon c_{\mathrm{BNN}} \left[b_\pi^{\mathrm{evt}}(\eta) A_s(\eta) - A_{s,\pi}(\eta) \sum_{\pi'} b_{\pi'}^{\mathrm{evt}}(\eta) \right].$$

Thus $b^{\mathrm{evt}} = 0$ makes the entire revision term vanish. Since Φ is continuous on the compact policy simplex and is nondecreasing along the trajectory, it has a finite limit Φ_∞. Fix $\eta^\infty \in \omega(\eta)$, choose $t_n \geq n$ with $\eta(t_n) \to \eta^\infty$, and set $\eta_n(t) = \eta(t_n + t)$ for $|t| \leq n$. The derivative bound and Lemma 17(i)–(ii) give, after relabeling a subsequence, local uniform convergence to a complete solution ζ of the same autonomous equation, with $\zeta(0) = \eta^\infty$. For each fixed $t \in \mathbb{R}$, continuity and $t_n + t \to \infty$ give

$$\Phi\left(\left(\sum_s \zeta_{s,\pi}(t) \right)_{\pi \in \Pi} \right) = \lim_{n\to\infty} \Phi(\nu(t_n + t)) = \Phi_\infty.$$

The equality condition in (94) therefore places the marginal of $\zeta(t)$ in $\mathcal{E}_\Pi^*$ and gives $b^{\mathrm{evt}}(\zeta(t)) = 0$ at every time. Hence ζ solves the frozen state–policy dynamics, with one constant marginal $\nu^* \in \mathcal{E}_\Pi^*$. Corollary 31 implies $\eta^\infty = \zeta(0) = \bar{\eta}^{\nu^*}$. Thus $\omega(\eta) \subseteq \mathcal{E}_{\mathcal{S}\times\Pi}^*$; compactness and continuity of the marginal map prove both limits in (69).

At a stationary BNN population, the potential derivative is zero, so (94) gives $\nu \in \mathcal{E}_\Pi^*$ and $b^{\mathrm{evt}} = 0$. The population is therefore frozen and stationary, hence equals $\bar{\eta}^\nu$ by Corollary 31. Conversely, every $\bar{\eta}^{\nu^*}$ with $\nu^* \in \mathcal{E}_\Pi^*$ has zero frozen drift and zero BNN revision. Theorem 9 identifies the aggregate populations in $\mathcal{E}_{\mathcal{S}\times\Pi}^*$ with $\mathcal{E}_{\mathcal{I}}^*$ and gives their common completion rates and reward. Uniform continuity on the compact state–policy simplex now gives (71).

(ii) Replicator revision. Equation (90) and Corollary 26(ii) give the policy limit ν^∞ in (70). At every $\eta \in \mathcal{E}_{\nu^\infty}$,

$$X_{s,\pi} p_{s,\pi\to\pi'}^{\mathrm{evt}} = \frac{c_{\mathrm{rep,evt}}}{m} \eta_{s,\pi} \nu_{\pi'}^\infty [H_{\pi'}(\nu^\infty) - H_\pi(\nu^\infty)]^+ = 0$$

by (72). This identity concerns the switching flux: a conditional probability from an absent source policy need not vanish, but the state–policy completion rate $X_{s,\pi}$ is zero. Every incoming and outgoing revision term in (89) contains one of these zero fluxes. Equation (89) therefore reduces on $\mathcal{E}_{\nu^\infty}$ to the frozen state–policy dynamics (51). Lemma 32, with $Z = \{\nu^\infty\}$, gives $\eta(t) \to \bar{\eta}^{\nu^\infty}$. If $\mathcal{E}_0(\nu) \cap \mathcal{E}_\Pi^* \neq \varnothing$, the maximum of Φ on this face equals its global maximum, so $\nu^\infty \in \mathcal{E}_\Pi^*$.

The marginal equation in (90) makes stationarity of ν equivalent to equal payoffs on $\operatorname{supp} \nu$. The same flux calculation then removes every revision term, so a stationary state–policy population must equal $\bar{\eta}^\nu$ by Corollary 31. Conversely, every such population has zero frozen drift and zero switching flux, and is stationary.

Continuity of aggregation gives $\mu(\eta(t)) \to \bar{\mu}^{\nu^\infty}$. Under the initial-face condition, $\nu^\infty \in \mathcal{E}_\Pi^*$, so Theorem 9 and continuity of completion rates and rewards give (71). □

These results concern the specified revision rules. The BNN proof uses the completion-weighted average payoff $\bar{H}_{\mathrm{evt}}(\eta)$ throughout the trajectory. Without the source-rate correction, the same sampling and adoption rule has switching fluxes proportional to $\chi_\pi(\eta)\nu_{\pi'}[H_{\pi'} - H_\pi]^+$, and its policy marginal generally differs from ordinary replicator dynamics.

F State–policy dynamics with vanishing revision rate and probability

We now allow the exogenous Poisson revision rate and the event-driven revision probability to tend to zero along a trajectory. Replace κ in (14) by $\kappa(t)$, and replace ε in (17) by $\varepsilon(t)$, where t denotes physical time in both systems. Agents continue to use the stationary payoffs $H(\nu(t))$ and a common measurable best-response selection $\gamma(t) \in \Gamma(\nu(t))$. Vanishing revision allows the state–policy population to approach its associated stationary population, while an infinite cumulative amount of revision ensures policy convergence to equilibrium.

Theorem 40 (Convergence with vanishing revision schedules)**.** *Under Assumptions 1 and 4, let $\kappa : [0,\infty) \to (0,\infty)$ and $\varepsilon : [0,\infty) \to (0,1]$ be continuous and satisfy*

$$\kappa(t) \longrightarrow 0, \qquad \varepsilon(t) \longrightarrow 0, \qquad \int_0^\infty \kappa(t')\,dt' = \int_0^\infty \varepsilon(t')\,dt' = \infty. \tag{95}$$

Each of (14) *and* (17)*, with its respective time-dependent revision parameter, has a global absolutely continuous solution from every initial point in $\mathcal{M}_m(\mathcal{S}\times\Pi)$. Every such solution has the following properties.*

(i) For state–policy dynamics with exogenous revision,

$$\Delta(\nu(t)) \le \exp\left(-\int_0^t \kappa(t')\,dt'\right)\Delta(\nu(0)), \tag{96}$$

$$M_{\mathrm{BR}}(\nu(t)) \le \exp\left(-\int_0^t \kappa(t')\,dt'\right)M_{\mathrm{BR}}(\nu(0)). \tag{97}$$

For event-driven state–policy dynamics,

$$\Delta(\nu(t)) \le \exp\left(-\underline{\lambda}\int_0^t \varepsilon(t')\,dt'\right)\Delta(\nu(0)). \tag{98}$$

(ii) For both systems,

$$\|\eta(t) - \bar{\eta}^{\nu(t)}\| \longrightarrow 0, \qquad \mathrm{dist}(\eta(t), \mathcal{E}^*_{\mathcal{S}\times\Pi}) \longrightarrow 0. \tag{99}$$

Equivalently, $\omega(\eta) \subseteq \mathcal{E}^*_{\mathcal{S}\times\Pi}$. The aggregate state–action population approaches $\mathcal{E}^*_{\mathcal{I}}$, and

$$\lambda(\mu(\eta(t))) \longrightarrow (1/\tau_i^*)_{i\in\mathcal{I}}, \qquad \frac{r\cdot x(\mu(\eta(t)))}{m} \longrightarrow g^*.$$

If $\mathcal{E}^*_\Pi = \{\nu^*\}$, then $\eta(t) \to \bar{\eta}^{\nu^*}$.

Proof. The revision rate $\kappa(t)$ and revision probability $\varepsilon(t)$ are bounded because they are continuous and tend to zero as $t \to \infty$. For either system, write its right-hand side as $F(t,\eta,\gamma)$, where

$$F_{s,\pi}(t,\eta,\gamma) = \begin{cases} A_{s,\pi} - X_{s,\pi} + \kappa(t)\big(\gamma_\pi m_s/m - \eta_{s,\pi}\big), & \text{exogenous revision,} \\ (1-\varepsilon(t))A_{s,\pi} - X_{s,\pi} + \varepsilon(t)\gamma_\pi A_s/m, & \text{event-driven revision,} \end{cases}$$

with all population-dependent quantities evaluated at η. Introduce the augmented state $u = (t,\eta)$ in the closed, locally compact set $K = [0,\infty)\times\mathcal{M}_m(\mathcal{S}\times\Pi)$, and define its policy marginal by $\ell(u) = (\sum_s \eta_{s,\pi})_{\pi\in\Pi}$. We seek a solution of the autonomous controlled system

$$\dot{u} = (1, F(t,\eta,\gamma)), \qquad \gamma \in \Gamma(\ell(u)), \qquad u(0) = (0,\eta(0)).$$

The augmented field is continuous in u and affine in γ. As shown in the proof of Lemma 24, $u \mapsto \Gamma(\ell(u))$ is upper semicontinuous with nonempty compact convex values. For every $\gamma \in \Gamma(\ell(u))$,

the population components of the field sum to zero. At a boundary point with $\eta_{s,\pi} = 0$, they satisfy $F_{s,\pi} = A_{s,\pi} + \kappa(t)\gamma_\pi m_s/m \geq 0$ for exogenous revision and $F_{s,\pi} = (1-\varepsilon(t))A_{s,\pi} + \varepsilon(t)\gamma_\pi A_s/m \geq 0$ for event-driven revision. These inequalities give tangency to the state–policy simplex; the clock component 1 points into $[0,\infty)$ at $t = 0$. Hence $(1, F(t,\eta,\gamma)) \in \mathrm{Tan}_K(u)$. The viability theorem used in that lemma's proof (Aubin & Frankowska, 1985, Theorem 2.1) therefore gives a local absolutely continuous solution in K with a common measurable best-response selection. Its first coordinate starts at zero and has derivative one, so it equals physical time; its population component consequently solves the required time-dependent equation. Inflow and outflow bounds give, almost everywhere,

$$\|\dot\eta(t)\|_1 \leq \begin{cases} 2m\big(\overline{\lambda} + \sup_{t'\geq 0}\kappa(t')\big), & \text{exogenous revision,} \\ 2m\overline{\lambda}, & \text{event-driven revision.} \end{cases}$$

These bounds and compactness of the state–policy simplex permit continuation to every finite time, as in the proof of that lemma.

(i) Summing the exogenous equation over states gives $\dot\nu = \kappa(t)(\gamma - \nu)$. Define its cumulative slow clock by

$$\widetilde{t}(t) := \int_0^t \kappa(t')\,dt'.$$

It is strictly increasing and tends to infinity by (95). In this clock, the policy marginal satisfies the unit-rate best-response inclusion. Theorem 8 therefore gives (96)–(97). For the event-driven system, summing over states gives $\dot\nu_\pi = \varepsilon(t)(\gamma_\pi\Lambda/m - \chi_\pi)$. All quantities in the following calculation are evaluated at time t. Since $\sum_\pi \gamma_\pi = m$ and $\sum_\pi \chi_\pi = \Lambda$, $\sum_\pi \dot\nu_\pi = \varepsilon(t)(\Lambda - \Lambda) = 0$. Moreover, $\gamma \in \Gamma(\nu)$ assigns positive mass only to maximizing policies, so

$$\sum_\pi \gamma_\pi \log H_\pi = \left(\sum_\pi \gamma_\pi\right) \log \max_\sigma H_\sigma = m \log \max_\sigma H_\sigma.$$

Using $\partial\Phi/\partial\nu_\pi = \log H_\pi - 1$, the chain rule gives, almost everywhere,

$$\begin{aligned} \frac{d}{dt}\Phi(\nu(t)) &= \sum_\pi (\log H_\pi - 1)\dot\nu_\pi = \sum_\pi \log H_\pi\, \dot\nu_\pi \\ &= \varepsilon(t)\left[\frac{\Lambda}{m}\sum_\pi \gamma_\pi \log H_\pi - \sum_\pi \chi_\pi \log H_\pi\right] \\ &= \varepsilon(t)\left[\Lambda \log \max_\sigma H_\sigma - \sum_\pi \chi_\pi \log H_\pi\right] \\ &= \varepsilon(t)\sum_\pi \chi_\pi \log \frac{\max_\sigma H_\sigma}{H_\pi}. \end{aligned}$$

For each policy, the completion-rate lower bound gives

$$\chi_\pi = \sum_s \eta_{s,\pi}\lambda_{s,\pi(s)}(\mu) \geq \underline{\lambda}\sum_s \eta_{s,\pi} = \underline{\lambda}\nu_\pi.$$

Since $\max_\sigma H_\sigma \geq H_\pi > 0$, every logarithmic ratio is nonnegative. Therefore,

$$\begin{aligned} \frac{d}{dt}\Phi(\nu(t)) &\geq \varepsilon(t)\underline{\lambda}\sum_\pi \nu_\pi \log \frac{\max_\sigma H_\sigma}{H_\pi} \\ &= \varepsilon(t)\underline{\lambda}\left[m \log \max_\sigma H_\sigma - \sum_\pi \nu_\pi \log H_\pi\right] \\ &= \varepsilon(t)\underline{\lambda} M_{\mathrm{BR}}(\nu(t)) \geq \varepsilon(t)\underline{\lambda}\Delta(\nu(t)), \end{aligned}$$

where the two equalities use $\sum_\pi \nu_\pi = m$ and the definition of M_{BR}, respectively, and the last inequality is the concavity bound (38). Since $\Delta = \Phi^* - \Phi$, it follows that

$$\frac{d}{dt}\left[e^{\underline{\lambda}\int_0^t \varepsilon(t')\,dt'}\Delta(\nu(t))\right] \leq 0$$

almost everywhere. Integrating from 0 to t proves (98). In either system, the divergent integral gives $\Delta(\nu(t)) \to 0$. Since Δ is continuous on the compact policy simplex and vanishes exactly on $\mathcal{E}^*_\Pi$, it follows that $\operatorname{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$.

(ii) Relative to the frozen field F_{phys} in (51), the revision terms satisfy, almost everywhere,

$$\|\dot\eta(t) - F_{\mathrm{phys}}(\eta(t))\|_1 \le \begin{cases} 2m\kappa(t), & \text{exogenous revision,} \\ 2m\overline{\lambda}\varepsilon(t), & \text{event-driven revision.} \end{cases} \tag{100}$$

The first bound follows by summing $\kappa(t)(\gamma_\pi m_s/m - \eta_{s,\pi})$ in absolute value. For the second, the physical-time revision term is $\varepsilon(t)\Lambda(\eta)F^{\mathrm{evt}}_{\mathrm{rev}}(\eta,\gamma)$; use $\|F^{\mathrm{evt}}_{\mathrm{rev}}\|_1 \le 2$ and $\Lambda \le m\overline{\lambda}$, as in (81).

Fix $\eta^\infty \in \omega(\eta)$, choose $t_n \ge n$ such that $\eta(t_n) \to \eta^\infty$, and set $\zeta_n(t) := \eta(t_n + t)$ for $|t| \le n$. The shifted trajectories lie in the compact state–policy simplex and have the common derivative bound above. For every fixed $T > 0$ and $n \ge T$, (100) gives

$$\sup_{a,b\in[-T,T]} \left\| \zeta_n(b) - \zeta_n(a) - \int_a^b F_{\mathrm{phys}}(\zeta_n(t'))\,dt' \right\|_1 \le \begin{cases} 4mT \sup_{|t'|\le T} \kappa(t_n + t'), & \text{exogenous revision,} \\ 4m\overline{\lambda}T \sup_{|t'|\le T} \varepsilon(t_n + t'), & \text{event-driven revision,} \end{cases} \longrightarrow 0.$$

Lemma 17(i)–(ii), with $I = \mathbb{R}$, $K = \mathcal{M}_m(\mathcal{S}\times\Pi)$, $T_n = n$, and $f = F_{\mathrm{phys}}$, gives a subsequence ζ_{n_k} and a Lipschitz map $\zeta : \mathbb{R} \to K$ such that

$$\sup_{|t|\le T} \|\zeta_{n_k}(t) - \zeta(t)\| \longrightarrow 0 \quad \text{for every } T > 0, \qquad \zeta(b) - \zeta(a) = \int_a^b F_{\mathrm{phys}}(\zeta(t'))\,dt' \quad (a, b \in \mathbb{R}).$$

Thus ζ is a complete frozen solution with $\zeta(0) = \eta^\infty$. Its policy marginal is constant and equals $\nu^\infty := \nu(\eta^\infty) \in \mathcal{E}^*_\Pi$, by part (i). Corollary 31 gives $\zeta \equiv \bar\eta^{\nu^\infty}$. We have therefore proved $\omega(\eta) \subseteq \mathcal{E}^*_{\mathcal{S}\times\Pi}$. Compactness gives the distance limit in (99). The map $\eta \mapsto \eta - \bar\eta^{\nu(\eta)}$ is continuous by (16) and Lemma 23, and vanishes at every omega-limit point. Compactness therefore gives $\|\eta(t) - \bar\eta^{\nu(t)}\| \to 0$, the first limit in (99).

Summing the state–policy masses over policies that prescribe action a in state s, as in (12), gives

$$\mu_{sa}(\eta(t)) - \bar\mu^{\nu(t)}_{sa} = \sum_{\pi:\pi(s)=a} \left(\eta_{s,\pi}(t) - \bar\eta^{\nu(t)}_{s,\pi}\right) \longrightarrow 0.$$

Thus $\|\mu(\eta(t)) - \bar\mu^{\nu(t)}\| \to 0$. Since $\operatorname{dist}(\nu(t), \mathcal{E}^*_\Pi) \to 0$, Theorem 9(iii) gives $\operatorname{dist}(\bar\mu^{\nu(t)}, \mathcal{E}^*_\mathcal{I}) \to 0$ and the stated completion-rate and reward limits for $\bar\mu^{\nu(t)}$. The triangle inequality gives

$$\operatorname{dist}(\mu(\eta(t)), \mathcal{E}^*_\mathcal{I}) \le \|\mu(\eta(t)) - \bar\mu^{\nu(t)}\| + \operatorname{dist}(\bar\mu^{\nu(t)}, \mathcal{E}^*_\mathcal{I}) \longrightarrow 0.$$

Uniform continuity of λ and x on the compact state–action simplex $\mathcal{M}_m(\mathcal{I})$ also gives

$$\|\lambda(\mu(\eta(t))) - \lambda(\bar\mu^{\nu(t)})\| \longrightarrow 0, \qquad \frac{\left|r \cdot [x(\mu(\eta(t))) - x(\bar\mu^{\nu(t)})]\right|}{m} \longrightarrow 0.$$

Combining these differences with the limits for $\bar\mu^{\nu(t)}$ yields $\lambda(\mu(\eta(t))) \to (1/\tau^*_i)_{i\in\mathcal{I}}$ and $r \cdot x(\mu(\eta(t)))/m \to g^*$. If the policy equilibrium set is a singleton, so is $\mathcal{E}^*_{\mathcal{S}\times\Pi}$, proving the final claim. □

Reciprocal revision schedules. A concrete choice satisfying (95) is

$$\kappa(t) = \frac{\kappa_0}{1 + t/t_0}, \qquad \varepsilon(t) = \frac{\varepsilon_0}{1 + t/t_0}, \qquad \kappa_0, t_0 > 0, \quad 0 < \varepsilon_0 \le 1, \tag{101}$$

where κ_0 is the initial Poisson revision rate, ε_0 is the initial revision probability, and t_0 is a physical-time scale. Their integrals are $\kappa_0 t_0 \log(1 + t/t_0)$ and $\varepsilon_0 t_0 \log(1 + t/t_0)$, respectively. Hence

$$\Delta(\nu(t)) \le \Delta(\nu(0)) \begin{cases} (1 + t/t_0)^{-\kappa_0 t_0}, & \text{exogenous revision,} \\ (1 + t/t_0)^{-\underline{\lambda}\varepsilon_0 t_0}, & \text{event-driven revision.} \end{cases}$$

The exogenous bound also holds with M_{BR} in place of Δ. Both state–policy trajectories approach $\mathcal{E}^*_{\mathcal{S}\times\Pi}$ by Theorem 40.

G Numerical Methods and Supplementary Results

G.1 Data and service area

We use the NYC TLC high-volume for-hire vehicle records for March 2026, with pickups between 9 and 10 a.m. on weekdays (New York City Taxi and Limousine Commission, 2026). Manhattan TLC data have also been used to study fleet control and vehicle relocation (Kullman et al., 2022; Yuan et al., 2022). Our model has a fixed driver population circulating within Manhattan, so we select trips with both endpoints there, including trips within a single zone. The sample covers 66 of the 69 Manhattan TLC IDs; IDs 103–105 (Governor's, Ellis, and Liberty Islands) have no trips in the filtered sample. We use these 66 zones and their official boundaries in every experiment (Figure 2).

We pool records from all 22 weekdays and exclude trips flagged as shared requests, shared matches, or Access-A-Ride trips, as well as records with driver pay outside $(0, 500]$ dollars, passenger-trip durations outside 1–120 minutes, or distances outside 0.05–100 miles. We also exclude records whose timestamp duration differs from the reported trip time by more than 60 seconds. Rewards use the TLC field `driver_pay`, which excludes tolls and tips and is net of commission, surcharges, and taxes (New York City Taxi and Limousine Commission, 2025). After filtering, the sample contains 217,958 trips across 3,824 origin–destination (OD) pairs, or 87.79% of the 66^2 pairs. Each zone has at least 21 pickups. Trips ending in Manhattan account for 79.38% of the quality-filtered trips starting there. All calibrated parameters are held fixed throughout each numerical experiment.

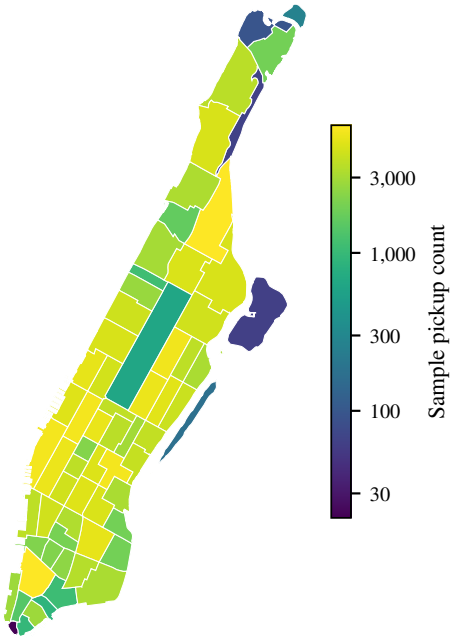


Figure 2: The 66 Manhattan TLC zones and their pickup counts in the March weekday-morning sample (logarithmic color scale). Map drawn from official NYC TLC zone polygons and trip records (New York City Taxi and Limousine Commission, 2026).

G.2 RIDE-HAILING MODEL

Following the ride-hailing setting of Courcoubetis & Dimakis (2025), we model drivers' strategic repositioning behavior, i.e., choosing where to seek their next passenger. Set $\mathcal{S} = \mathcal{A} = \{1, \dots, 66\}$. A driver starting a service cycle in zone s chooses pickup zone a, relocates there, and serves a passenger whose destination is s'. The cycle ends in state s' and pays reward r_{sa}.

Let $n_{ss'}^{\mathrm{trip}}$ be the number of observed trips from s to s', with $n_s^{\mathrm{trip}} = \sum_{s'} n_{ss'}^{\mathrm{trip}}$ and $n^{\mathrm{trip}} = \sum_s n_s^{\mathrm{trip}}$. We smooth the pickup and drop-off shares by adding one to each zone count:

$$p_s^{\mathrm{pick}} = \frac{n_s^{\mathrm{trip}} + 1}{n^{\mathrm{trip}} + |\mathcal{S}|}, \qquad p_{s'}^{\mathrm{drop}} = \frac{\sum_s n_{ss'}^{\mathrm{trip}} + 1}{n^{\mathrm{trip}} + |\mathcal{S}|}.$$

For the destination probability distribution conditional on pickup zone a, we use p^{drop} as a prior with total weight one:

$$P_{ss'}^a = \frac{n_{as'}^{\mathrm{trip}} + p_{s'}^{\mathrm{drop}}}{n_a^{\mathrm{trip}} + 1}, \qquad s, s' \in \mathcal{S}, \quad a \in \mathcal{A}.$$

All 66 pickup actions are available in every state. Each destination has observed arrivals from at least one pickup zone, so the unsmoothed model already satisfies Assumption 4: from any state, a driver can choose a pickup zone with a positive probability of reaching the desired destination. Smoothing makes every transition probability positive and every deterministic policy irreducible.

We normalize the population mass to $m = 1$ and measure physical time in minutes. Let t_a^{ride} and r_{sa} be the mean passenger-trip duration and driver pay from pickup zone a, respectively, each smoothed toward its sample-wide mean with prior weight one. Rewards depend only on a. The sample means are $\bar{t}^{\mathrm{ride}} = 16.68$ minutes and \$16.15; the smoothed rewards range from \$13.44 to \$40.07 and are strictly positive.

We estimate empty relocation times from observed passenger-trip times. For $s \neq a$, let t_{sa}^{total} be total passenger-trip minutes from s to a, and let d_{sa}^{geo} be the straight-line distance between zone centroids in kilometres. We compute each zone's polygon centroid from the official NYC TLC taxi-zone boundary shapefile (New York City Taxi and Limousine Commission, 2026), using its NAD83 / New York Long Island projected coordinate system. We then take Euclidean distances between these centroids and convert from US survey feet to kilometres. We smooth the OD travel-time estimates using a distance-based prior with weight 20:

$$t_{sa}^{\mathrm{move}} = \frac{t_{sa}^{\mathrm{total}} + 20 \max\{1, b_{\mathrm{fit},0} + b_{\mathrm{fit},1} d_{sa}^{\mathrm{geo}}\}}{n_{sa}^{\mathrm{trip}} + 20} \quad (s \neq a), \qquad t_{ss}^{\mathrm{move}} = 0.$$

The coefficients $b_{\mathrm{fit},0} = 9.231$ minutes and $b_{\mathrm{fit},1} = 2.650$ minutes/km come from a nonnegative least-squares fit of OD mean duration to centroid distance, weighted by trip counts. The distance estimate supplies relocation times for unobserved OD pairs and moderates estimates based on few trips.

Driver idle times are not observed. We specify a matching delay with a two-minute baseline and a congestion term that increases with the total completion rate at the pickup zone. Write

$$\tau_{sa}^{\mathrm{base}} = t_{sa}^{\mathrm{move}} + t_a^{\mathrm{ride}} + 2, \qquad \beta_a = \frac{5\bar{t}^{\mathrm{ride}}}{p_a^{\mathrm{pick}}}.$$

The sojourn times and potential are

$$\tau_{sa}(x) = \tau_{sa}^{\mathrm{base}} + \beta_a \sum_{s'} x_{s'a}, \qquad G(x) = \sum_{s,a} \tau_{sa}^{\mathrm{base}} x_{sa} + \frac{1}{2} \sum_a \beta_a \Big(\sum_s x_{sa} \Big)^2.$$

The congestion term can be written as

$$\beta_a \sum_{s'} x_{s'a} = 5 \, \frac{\sum_{s'} x_{s'a}}{p_a^{\mathrm{pick}} / \bar{t}^{\mathrm{ride}}} \quad \text{minutes.}$$

With $m = 1$, we allocate the reference total completion rate $1/\bar{t}^{\mathrm{ride}}$ across zones in proportion to their observed pickup shares. This assumes that zones with greater pickup activity can accommodate

larger completion rates at the same added delay. The reference rate uses passenger-trip duration alone, excluding relocation and waiting times. The five-minute scale is an illustrative model parameter; the trip records do not provide matching-delay observations. This quadratic G satisfies Assumption 1: it is convex and smooth, and $\partial_{sa} G(x) = \tau_{sa}(x) > 0$ on $\mathbb{R}^{\mathcal{I}}_+$. We represent the service cycle by a single action with instantaneous completion rate $1/\tau_{sa}(x(\mu))$. Exogenous revision can change the pickup action during a cycle, as in (14); event-driven revision occurs after the current cycle is completed.

G.3 Exploitability and stationarity gap

The two quantities plotted in Figure 1 measure policy optimality and population stationarity, respectively. We define exploitability as the average gain from switching to a best response at the current policy distribution (Perrin et al., 2020):

$$\mathrm{Expl}(\nu) = \max_{\pi \in \Pi} H_\pi(\nu) - \frac{1}{m} \sum_{\pi \in \Pi} \nu_\pi H_\pi(\nu).$$

The payoffs $H_\pi(\nu)$ are evaluated at $\bar{\mu}^\nu$. Exploitability is measured in dollars per minute and vanishes exactly on $\mathcal{E}^*_\Pi$. The best-response gap M_{BR} in Theorem 8 uses logarithmic payoffs; exploitability expresses the improvement in the original payoff units.

For a state–policy population η with policy marginal ν, we define the frozen stationarity gap by

$$\mathrm{Sgap}(\eta) = \frac{\|F_{\mathrm{phys}}(\eta)\|_1}{\Lambda(\eta)} = \|F_{\mathrm{evt}}(\eta)\|_1 = \frac{1}{\Lambda(\eta)} \sum_{s,\pi} \left| \sum_{s'} X_{s',\pi}(\eta) P^{\pi(s')}_{s's} - X_{s,\pi}(\eta) \right|.$$

Here F_{phys} and $F_{\mathrm{evt}} = F_{\mathrm{phys}}/\Lambda$ are the frozen state–policy vector fields in physical time and event time, respectively. The gap is the total absolute imbalance between arrival and completion rates, normalized by the total completion rate. It measures the drift that remains when policy revision is switched off. Like exploitability, it has a maximization form:

$$\mathrm{Sgap}(\eta) = \max_{\|v\|_\infty \le 1} v \cdot F_{\mathrm{evt}}(\eta), \qquad v \in \mathbb{R}^{\mathcal{S} \times \Pi}.$$

It is therefore the largest instantaneous drift, per unit event time, of $\sum_{s,\pi} v_{s,\pi} \eta_{s,\pi}$ along the frozen dynamics, over test functions bounded by one in absolute value. It lies in $[0, 2]$: it is twice the total-variation distance between the normalized arrival and completion rates. By Proposition 10, $\mathrm{Sgap}(\eta) = 0$ exactly when $\eta = \bar{\eta}^\nu$. Consequently, the two gaps vanish together exactly on $\mathcal{E}^*_{\mathcal{S} \times \Pi}$.

G.4 Computation and experimental setup

At each update with policy revision, we first compute the stationary population $\bar{\mu}^\nu$ and the payoffs of policies currently used in the population. We solve the policy mass-balance equations in Section 4; for this model, this requires solving for the 66 pickup-zone congestion delays $\beta_a \sum_s \bar{x}^\nu_{sa}$.

We then solve the best-response LP (22) with sojourn times fixed at $\tau = \nabla G(\bar{x}^\nu)$, using the HiGHS simplex solver. The LP maximizes the total reward rate over nonnegative completion rates for all 66^2 state–action pairs, subject to flow balance and $\sum_{s,a} \tau_{sa} x_{sa} = m$. Its optimal value divided by m gives $\max_\pi H_\pi(\nu)$. The positive-flow action in each state of an optimal basic solution defines the best-response policy. The solver uses primal and dual feasibility tolerances of 10^{-10} and reuses the previous simplex basis. We store policies as they enter the population; every LP includes all actions.

We hold the selected best response fixed during the integration step. For the state–policy systems, the transition rates are evaluated at the current population, so we also solve Little's law (1) for $x(\mu)$. This calculation reduces to 66 independent scalar equations, one per pickup zone. Both this calculation and the stationary mass-balance calculation use a maximum absolute residual tolerance of 2×10^{-11}.

To set the time unit, we compute a reference equilibrium by solving the convex program (3) with Clarabel through CVXPY. We refine x^* through the active KKT equations and check optimality with the best-response LP at $\tau = \nabla G(x^*)$. The mean cycle time is $t_{\mathrm{ref}} = m / \sum_{s,a} x^*_{sa} = 24.63$ minutes.

All four runs in Figure 1 initially put all policy mass on the policy with the shortest estimated relocation time to another zone, $\arg\min_{a \neq s} t^{\mathrm{move}}_{sa}$, with ties broken by zone index. Staying in the

current zone is allowed; the restriction $a \neq s$ is used only to specify the initial policy. This policy stays fixed in panel (a), where the state–policy masses equal the corresponding state–action masses. Panels (a), (c), and (d) start with uniform state masses. Panel (b) uses $\bar{\eta}^{\nu}$ at every time, as prescribed by the policy dynamics. The initial exploitability is 0.316 dollars/minute in all panels; the initial stationarity gap is 0.556 in (a), (c), and (d), and zero in (b).

Panel (a) uses t/t_{ref}, and panel (b) uses the time of the unit-rate best-response policy dynamics. For exogenous revision (c), we use slow physical time $\widetilde{t} = \kappa t$; for event-driven revision (d), we use slow event time $\widetilde{\theta} = \varepsilon\theta$, where $\theta(t) = \int_0^t \Lambda(\eta(t'))\,dt'$. With $m = 1$, setting $\kappa t_{\mathrm{ref}} = \varepsilon$ matches the population-wide revision rates at the reference equilibrium. Figure 1 uses $\kappa t_{\mathrm{ref}} = \varepsilon = 0.01$. It shows $t/t_{\mathrm{ref}} \leq 50$ in (a), policy time up to 6 in (b), $\widetilde{t} \leq 6$ in (c), and $\widetilde{\theta} \leq 6$ in (d). The endpoints in (c) and (d) correspond to 10.26 and 10.76 days of physical time, respectively, with the morning calibration held fixed.

Best-response policy dynamics use the exact exponential update for a fixed best response. The state–action and both state–policy systems use positivity-preserving two-stage Runge–Kutta updates. The maximum step is 0.25 in t/t_{ref} for panel (a) and 2.5×10^{-4} in the respective clocks of panels (b)–(d). Completion outflows are limited by a stability factor of 0.025.

The reference equilibrium satisfies the mass, flow-balance, and LP optimality conditions to 10^{-10}. Along the trajectories, mass errors are below 10^{-12} and residuals in Little's law are below 2.10×10^{-11}. We repeat (b)–(d) with half the maximum step and all four runs with half the stability factor. At the saved observation times, exploitability changes by at most 4.30×10^{-4} dollars/minute and the stationarity gap by at most 7.90×10^{-5}. For the revision-frequency comparison, the same refinement at paired values $\kappa t_{\mathrm{ref}} = \varepsilon \in \{0.005, 0.01, 1\}$ changes the tabulated maxima by at most 1.30×10^{-5}. The reproduction files include the zone IDs, processed data, solver settings, and software versions.

G.5 Effect of revision frequency

Figure 3 varies κt_{ref} for exogenous revision and ε for event-driven revision over $\{0.01, 0.1, 1\}$, with the data and initial population fixed. The exploitability curves under exogenous revision nearly coincide: in slow physical time, the policy marginal satisfies $d\nu/d\widetilde{t} \in \Gamma(\nu) - \nu$, independently of κ. In the event-driven system, policy revision depends on completion shares, so the policy marginal remains coupled to the state–policy population. In both rows, increasing the revision frequency leaves a larger stationarity gap even after exploitability has fallen. Low policy exploitability can therefore coexist with substantial flow imbalance.

Table 1 reports the maximum stationarity gap over saved slow times in $[1, 6]$: $\widetilde{t}$ for exogenous revision and $\widetilde{\theta}$ for event-driven revision. It covers eight paired settings with $\kappa t_{\mathrm{ref}} = \varepsilon$. Reducing these values from 1 to 0.01 lowers the maximum by more than two orders of magnitude in both systems. Less frequent revision allows more population adjustment between policy changes, consistent with the rare-revision limits in Theorems 11 and 13. The comparison holds the slow-time window fixed; smaller revision parameters correspond to longer physical observation windows.

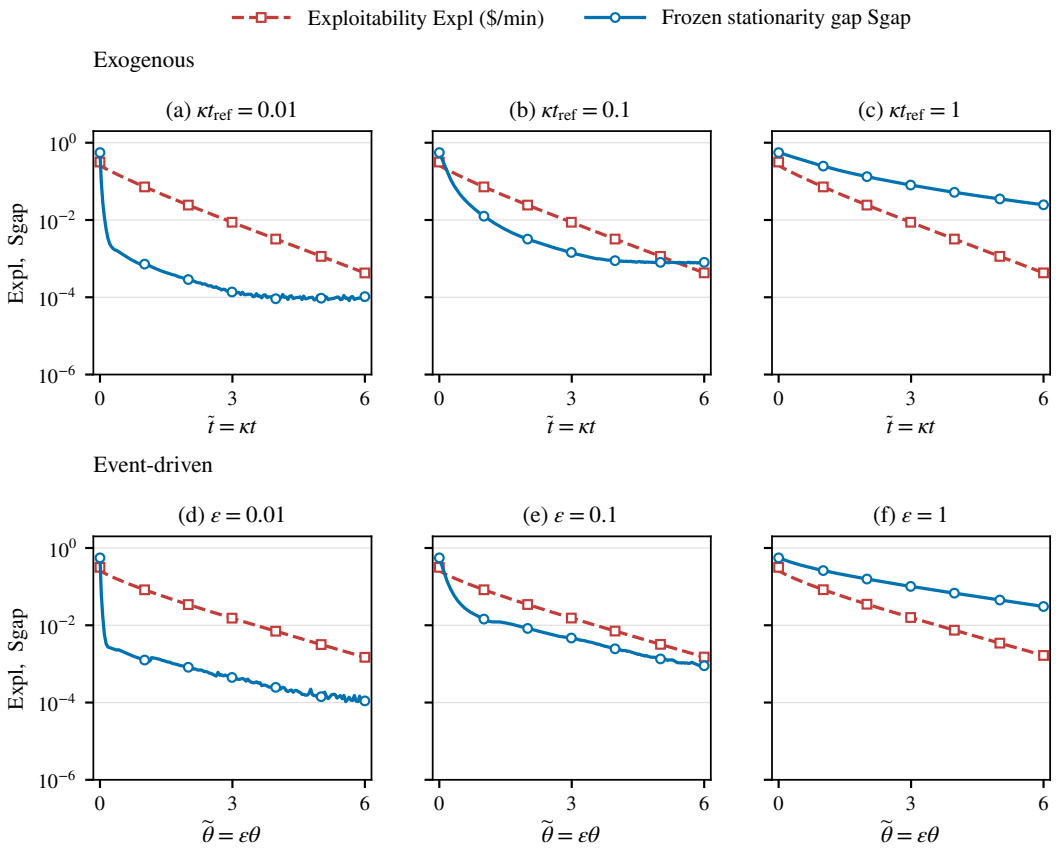


Figure 3: Exploitability and frozen stationarity gap under state–policy dynamics with exogenous revision (top) and event-driven state–policy dynamics (bottom). Columns pair the revision settings $\kappa t_{\mathrm{ref}} = \varepsilon \in \{0.01, 0.1, 1\}$.

Table 1: Maximum frozen stationarity gap Sgap over saved slow times in $[1, 6]$: $\tilde{t}$ for exogenous revision and $\tilde{\theta}$ for event-driven revision.

$\kappa t_{\mathrm{ref}} = \varepsilon$	0.005	0.01	0.02	0.05	0.1	0.2	0.5	1
Exogenous	3.72×10^{-4}	7.21×10^{-4}	1.42×10^{-3}	3.73×10^{-3}	0.013	0.041	0.136	0.248
Event-driven	7.54×10^{-4}	1.45×10^{-3}	2.83×10^{-3}	6.56×10^{-3}	0.014	0.038	0.138	0.261