STATIONARY ROBUST MEAN-FIELD GAMES UNDER MODEL MISMATCHES

Yue Wang
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, Florida, USA

June 23, 2026

ABSTRACT

Deploying multi-agent reinforcement learning (MARL) in the real world is often limited by model mismatches between the training simulators and the true environment, which could be further amplified through strategic interactions and result in severe performance degradation upon deployment. Distributional robustness offers a principled response by optimizing policies against worst-case transition models drawn from an uncertainty set, but standard robust MARL frameworks become increasingly intractable as the number of agents grows. This paper develops an infinite-horizon, stationary mean-field game framework that incorporates distributional model uncertainty directly into the population-coupled dynamics. We establish a robust dynamic programming principle with a contractive Bellman operator and prove the existence of a stationary robust mean-field equilibrium via a fixed-point argument. We further develop the first concrete algorithm with convergence guarantees. We then connect the mean-field solution to a finite-population robust game whose ambiguity sets depend on the empirical distribution, showing that the mean-field equilibrium policy induces approximate equilibrium behavior as the population size increases. Under a contractive robust-dynamics regime, we further obtain explicit non-asymptotic error bounds. Numerical experiments further illustrate the qualitative and quantitative impact of robustness under multiple uncertainty models, validating our theoretical findings.

1 Introduction

Multi-agent reinforcement learning (MARL), together with its Markov/stochastic-game formulation [Shapley, 1953, Littman, 1994], has become a core paradigm for designing intelligent multi-agent systems (MAS) that exhibit complex and coordinated behavior. By enabling multiple decision-makers to learn and act in a shared, evolving environment, MARL has achieved great success in, e.g., strategic games [Silver et al., 2017, Vinyals et al., 2019], coordination in autonomous transportation and traffic systems [Shalev-Shwartz et al., 2016, Hua et al., 2025], and distributed robotics [Lowe et al., 2017, Matignon et al., 2012].

Despite this progress, a fundamental obstacle that limits the reliable deployment of MARL in the real world is the *Sim-to-Real* gap [Zhao et al., 2020, Peng et al., 2018]. The standard workflow trains policies in a simulator and then deploys them in practice; yet, even high-fidelity simulators inevitably miss aspects of reality, including subtle physical effects, sensor imperfections, unmodeled dynamics, and latent environmental factors [Padakandla et al., 2020, Rajeswaran et al., 2016]. As a consequence, policies that appear optimal in simulation can be brittle under model mismatch and may degrade severely (or fail catastrophically) upon deployment in practice [Pinto et al., 2017].

This challenge becomes even more acute in multi-agent systems, as the model mismatches can be further endogenously amplified by interactions among agents. A small deviation in one agent's realized dynamics can change its behavior; and this further modifies the effective environment faced by others, prompting responses that further alter the joint dynamics. Such feedback loops can induce strong non-stationarity beyond that caused by strategic adaptation alone [Papoudakis et al., 2019, Canese et al., 2021, Wong et al., 2023]. This amplification effect makes the system-level

outcome of MARL to be highly sensitive to small modeling errors, and makes robustness against model mismatches a crucial concern.

A principled strategy to address model mismatches is *distributional robustness*. Rather than optimizing for a single nominal model, (distributionally) robust MARL optimize against a family of plausible transition models (an *uncertainty set*), selecting policies that maximize worst-case performance among them [Zhang et al., 2020, Kardeş et al., 2011]. This minimax viewpoint provides formal performance guarantees whenever the true environment lies in the uncertainty set, and it often acts as a regularizer that improves generalization under perturbations [Abdullah et al., 2019, Vinitsky et al., 2020, Liu et al., 2025]. However, existing results for distributionally robust MARL or Markov games suffer from the *curse of multi-agency*, which is a fundamental obstruction that the learning complexity scales exponentially in the number of agents [Farhat et al., 2026, Shi et al., 2024]. Thus, robust RL can become significantly inefficient and infeasible in large-scale multi-agent systems, where robustness is most needed.

On the other hand, Mean-field games (MFGs) [Huang et al., 2006, Lasry and Lions, 2007] offer a classical lens for large-population strategic interactions by exploiting symmetry and weak coupling. Instead of tracking the full N-agent joint state, MFGs approximate agent-wise interactions through the population distribution (the *mean field*), enabling scalable equilibrium computation and, under appropriate conditions, finite-N approximation guarantees [Carmona and Delarue, 2013, Saldi et al., 2018, Anahtarci et al., 2023, Cui and Koeppl, 2021, Yardim et al., 2023, 2024].

Given the effectiveness and efficiency of MFGs in large-scale MAS, a natural question arises: *Can we utilize the mean-field approximation to reduce the complexity of* stationary *(time-homogeneous), infinite-horizon distributionally robust Markov games, making robust MARL tractable under large systems?*

We answer this question affirmatively by developing a framework of discounted stationary distributionally robust mean-field games. Our contributions are summarized as follows.

Formulation and Solvability of robust MFGs with infinite-horizon discounted reward. We first propose the formulation of infinite-horizon robust MFGs under model mismatches, and their solution notions: robust mean-field equilibrium (MFE). We then prove the existence of a robust MFE under a standard assumption, confirming the solvability of discounted stationary robust MFGs. Different from existing studies on finite-horizon robust MFGs [Langner et al., 2024] which rely on backward construction of non-stationary robust MFE, our study relies on a fixed-point argument to handle the stationarity.

Convergent algorithms for robust MFE. We then design a concrete algorithm to identify the stationary robust MFE. We first develop a fixed-point-based characterization of the robust MFE, which enables us to obtain a robust MFE by finding the fixed point of an operator. We then develop a Robust Best-Response Iteration algorithm, and prove that it converges to a fixed point (and hence finds a robust MFE) under a standard assumption. Our results represent the first convergent algorithms under robust MFGs, providing a concrete algorithmic solution to uncertain MAS.

Approximation of finite-player robust games. We further develop comprehensive studies on the connections between robust MFGs and finite-player robust games (as the player number becomes large), to understand the effectiveness of approximating large-scale MAS through robust MFGs. We first reveal the hardness of such an approximation under the robust setting, and, unlike the non-robust cases, the necessity of additional stabilizing assumptions. We then develop an asymptotic approximation of robust MFG under this additional assumption, and further characterize the non-asymptotic approximation rate under additional quantitative assumptions. Our results hence enable tractable robust MARL under large population.

2 Related Work

Mean-field games. The framework of standard (non-robust) mean-field games (i.e., without uncertainty) are proposed and studied in both continuous-time (see, e.g., [Huang et al., 2006, Tembine et al., 2013, Huang, 2010, Gomes et al., 2013, Lacker, 2016, Lacker and Soret, 2022, Lacker and Zariphopoulou, 2019, Aurell et al., 2022, Delarue et al., 2020]) and in discrete-time (see, e.g., [Adlakha et al., 2015, Biswas, 2015, Gomes et al., 2010, Gast et al., 2012, Elliott et al., 2013, Moon and Başar, 2015, 2016a, Nourian and Nair, 2013, Saldi et al., 2018, 2019, Elie et al., 2020]). We also refer to [Carmona et al., 2018, Bensoussan et al., 2013, Gomes and Saúde, 2014, Laurière et al., 2022] for survey papers including both settings.

Under the regime of reinforcement learning in MFGs, extensive studies on algorithm design and convergence analysis are developed in, e.g., [Carmona and Delarue, 2013, Saldi et al., 2020, Anahtarci et al., 2023, Cui and Koeppl, 2021, Yardim et al., 2023, 2024, Perrin et al., 2020, Guo et al., 2019, Xie et al., 2021].

However, all of these existing works are for nominal mean-field games, and no model uncertainty is considered.

Distributionally robust Markov decision processes and Markov games. To address the model mismatch and the Sim-to-Real gap, distributionally robust Markov decision processes are first proposed and studied in [Iyengar, 2005, Nilim and El Ghaoui, 2004, Wiesemann et al., 2013]. Based on them, robust single-agent RL has been extensively studied under different settings, e.g., online learning [Wang and Zou, 2021, Lu et al., 2024, He et al., 2025, Ghosh et al., 2026, 2025], offline learning [Shi and Chi, 2024, Blanchet et al., 2023, Wang et al., 2024c,a], or with a generative model [Liu et al., 2022, Panaganti and Kalathil, 2022, Yang et al., 2022, Shi et al., 2023, Wang et al., 2024b, Xu et al., 2023, Roch et al., 2025b,a, Wang et al., 2024d, 2023c].

It is then further extended to the multi-agent regime, where the framework of distributionally robust Markov games is developed and studied firstly in [Kardeş et al., 2011, Zhang et al., 2020]. However, designing efficient MARL algorithms for distributionally robust Markov games is significantly challenging. Since learning the worst-case performance requires a global knowledge of the uncertainty, recent research reveals an inherent curse of multi-agency in robust MARL: the sample complexity exponentially depends on the number of agents [Shi et al., 2024, Blanchet et al., 2023, Farhat et al., 2026], unless additional assumptions or oracles are assumed [Shi et al., 2025, Jiao and Li, 2024, Ma et al., 2023, Li et al., 2026]. This inherent challenge prevents the deployment of robust MARL in large multi-agent systems, motivating other formulations and efficient solutions.

Mean-field games under model mismatches. When model mismatches are considered in the MFG framework, a promising approach is to consider robust MFGs. The most related works to ours are [Langner et al., 2024, Liang et al., 2026]. In [Langner et al., 2024], the framework of robust MFG is considered under the finite-horizon settings, and the existence of a non-stationary MFE is proved. However, their proofs are based on a backward construction argument, which is infeasible under our settings, and we develop a fundamentally different proof technique to handle the infinite-horizon dependence in our problems. Another most relevant work is recently developed in [Liang et al., 2026], where stationary robust MFGs are considered. However, this work studies a different formulation of the MFE, and only derives its existence under some strong conditions. Hence, although our studies bear similarities to them, our problem formulation and proof technique are fundamentally different. Moreover, neither of these works studies algorithm convergence or non-asymptotic approximation of robust MFGs.

Robust MFGs are also studied in continuous-time settings, e.g., [Moon and Başar, 2016b,c, Huang and Huang, 2013, Bauso et al., 2016]. However, the studies cannot be applied to our discrete settings, and do not subsume our studies. Another line of research studies robust mean-field control problems [Zaman et al., 2024, Laurière et al., 2025, Xu et al., 2025], which can be viewed as a cooperative MFG and shares fundamentally different objectives with ours.

3 Preliminaries

In massive MAS, computing exact equilibria via traditional game-theoretic methods becomes computationally intractable due to the exponential growth of the joint state-action space. Mean Field Games (MFGs) circumvent this curse by considering the asymptotic regime where the number of agents approaches infinity. In this limit, agents become infinitesimal, anonymous, and indistinguishable. Consequently, direct agent-to-agent interactions are replaced by the interaction between a single representative agent and the macroscopic state distribution of the entire population, termed the mean field. By tracking this aggregate distribution rather than individual microscopic states, the intractable N-player game elegantly decouples into a tractable local optimization problem coupled with a global consistency condition.

A stationary MFG [Anahtarci et al., 2023, Yardim et al., 2023, Guo et al., 2019, Xie et al., 2021] can be formulated as a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}, \mathcal{A}$ are the state and action spaces, with transition dynamics $P : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ and rewards $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \Delta(\mathcal{S}) \to \mathbb{R}$. $\gamma \in (0, 1)$ is the discount factor.

As mentioned, performances in a MFG are determined by the policy of the representative agent and the overall strategy of all other agents. Specifically, these strategies are characterized as a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and a population distribution $\mu \in \Delta(\mathcal{S})$. For any $(\mu, \pi) \in \Delta_{\mathcal{S}} \times \Pi$, the discounted infinite horizon value function is defined as

$$J_\mu(\pi, P) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu) \middle| \begin{smallmatrix} s_0 \sim \mu, a_t \sim \pi(s_t) \\ s_{t+1} \sim P(s_t, a_t, \mu) \end{smallmatrix}\right].$$

The goal of a MFG is to find some equilibrium:

Definition 1. *A policy-population pair $(\mu^\star, \pi^\star) \in \Delta_{\mathcal{S}} \times \Pi$ is called a Nash-equilibrium (or mean field equilibrium) if the two conditions hold:*

$$\textit{Stability:} \quad \mu^\star(s') = \sum_{s,a} \mu^\star(s)\pi^\star(a|s)P(s'|s,a,\mu^\star),$$
$$\textit{Optimality:} \quad J_{\mu^\star}(\pi^\star, P) = \max_{\pi\in\Pi} J_{\mu^\star}(\pi, P).$$

4 Infinite-horizon stationary robust mean-field games

We then introduce the major objective of our studies, the stationary (time-homogeneous) robust mean-field games. Similarly to a standard MFG, a robust MFG is specified as $(\mathcal{S}, \mathcal{A}, \mathfrak{P}, r)$ [Langner et al., 2024, Liang et al., 2026], where $\mathcal{S}, \mathcal{A}$ are the finite state and action spaces and r is the reward function. The potential Sim-to-Real gap is modeled through the model uncertainty sets (independently) defined through a set-valued map as

$$\begin{aligned} \mathfrak{P} :& \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightrightarrows \Delta(\mathcal{S}), \\ & (s, a, \mu) \mapsto \mathfrak{P}(s, a, \mu) \subseteq \Delta(\mathcal{S}), \end{aligned} \tag{1}$$

where each element of $\mathfrak{P}(s, a, \mu)$ is a candidate transition law for the next state from current (s, a) and population μ.

Under the model mismatches $\mathfrak{P}$, robust learning takes a principle of pessimism, and considers the worst-case performance over the uncertainty set $\mathfrak{P}$. For a pair (π, μ), the robust value function is defined as:

$$V_\mu^\pi := \inf_{p\in\mathfrak{P}(\mu)} J_\mu(\pi, p), V_\mu := \sup_\pi \inf_{p\in\mathfrak{P}(\mu)} J_\mu(\pi, p), \tag{2}$$

where $\mathfrak{P}(\mu) \triangleq \otimes_{(s,a)}\mathfrak{P}(s, a, \mu)$ and the infimum takes over kernels with $p(\cdot \mid s, a) \in \mathfrak{P}(s, a, \mu)$ for any (s, a).

We then extend the definition of MFE and introduce the notion of robust equilibrium as follows.

Definition 2 (Stationary robust mean-field equilibrium)**.** *A triple $(\mu^\star, \pi^\star, p^\star)$ with $\mu^\star \in \Delta(\mathcal{S})$, $\pi^\star : \mathcal{S} \to \Delta(\mathcal{A})$, and $p^\star : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ is a stationary robust mean-field equilibrium if:*[1]

(i) ***Robust optimality:*** *$\pi^\star$ attains* (2) *at $\mu^\star$:*

$$V_{\mu^\star} = \inf_p J_{\mu^\star}(\pi^\star, p). \tag{3}$$

(ii) ***Worst-case kernel:*** *$p^\star$ attains the infimum against $\pi^\star$:*

$$V_{\mu^\star} = J_{\mu^\star}(\pi^\star, p^\star). \tag{4}$$

(iii) ***Consistency (stationarity):*** *$\mu^\star$ is invariant under $(\pi^\star, p^\star)$: for any $s' \in \mathcal{S}$:*

$$\mu^\star(s') = \sum_{s\in\mathcal{S}} \mu^\star(s) \sum_{a\in\mathcal{A}} \pi^\star(a|s)\, p^\star(s'|s, a, \mu^\star). \tag{5}$$

Specifically, if we view the three components as three players: a representative player (who controls π), the population player (controls μ), and an environment player (controls p), a robust MFE is a triple where the three players achieve a balanced state. Such a notion is an extension of the Nash equilibrium of standard non-robust MFGs to the worst-case [Huang et al., 2006, Lasry and Lions, 2007]. Moreover, a similar notion of robust MFE is studied in [Langner et al., 2024] for the non-stationary finite-horizon setting.

5 Solvability of Robust MFGs

In this section, we study the solvability of infinite-horizon robust MFGs, i.e., if the stationary robust MFE in Definition 2 always exists.

We first make the following standard assumptions.

Assumption 1. *We assume the following conditions hold: (1). $\mathcal{S}$ and $\mathcal{A}$ are finite; (2). For every (s, a, μ), the set $\mathfrak{P}(s, a, \mu)$ is nonempty, convex, and compact in $\Delta(\mathcal{S})$. Moreover, for every fixed (s, a) the correspondence $\mu \mapsto \mathfrak{P}(s, a, \mu)$ is continuous:*

[1] In Proposition 3, we showed that the worst-case can be achieved by stationary kernel, instead of history-dependent ones.

(a) ***(Closed graph)*** *if $\mu_n \to \mu$, $p_n \to p$, and $p_n \in \mathfrak{P}(s, a, \mu_n)$ for all n, then $p \in \mathfrak{P}(s, a, \mu)$.*

(b) ***(Lower hemicontinuity)*** *if $\mu_n \to \mu$ and $p \in \mathfrak{P}(s, a, \mu)$, then there exist $p_n \in \mathfrak{P}(s, a, \mu_n)$ such that $p_n \to p$.*

And (3). r is bounded and continuous in μ (w.r.t. the Euclidean topology on $\Delta(\mathcal{S})$).

Remark 1. *Parts (1) and (3) are standard in standard MFG studies [Huang et al., 2006]; Part (2) is adapted from the continuity assumption on the transition kernel in standard MFGs and standard robust RL literature [Iyengar, 2005, Wang et al., 2023b, Wang and Si, 2025], and can be easily satisfied by many standard uncertainty sets, e.g., distributionally ambiguous sets defined by divergence.*

Under these standard assumptions, we then derive the existence result as follows.

Theorem 1 (Existence of a stationary robust mean-field equilibrium)**.** *Under Assumption 1, there exists a stationary robust mean-field equilibrium $(\mu^\star, \pi^\star, p^\star)$.*

Our result hence implies that, for an infinite horizon robust MFG with discounted reward, there always exists a mean-field equilibrium, and the game is always solvable. We also highlight that our proof is fundamentally different from the finite-horizon case [Langner et al., 2024], where the non-stationary mean-field equilibrium can be constructed through a backward induction, whereas in our setting, the existence of a stationary equilibrium is derived through fixed-point arguments.

6 Approximation of Finite-Player Robust Games

In this section, we aim to show that, under some assumptions, the policy $\pi^\star$ of a robust MFE constitutes an approximate Nash equilibrium of the finite-player robust Markov game (defined below). We study both asymptotic and non-asymptotic convergence, developing a comprehensive understanding of the connections.

6.1 N-player robust games

In this section, we first introduce the finite N-player robust game. Fix $N \in \mathbb{N}$, a N-player robust game is specified as $(N, \mathcal{S}, \mathcal{A}, \mathfrak{P}, r, \mu^\star)$. For each agent $i \in N = \{1, \dots, n\}$, it has $s_t^i \in \mathcal{S}$ and $a_t^i \in \mathcal{A}$ as its local (individual) state/action, with $s_0^i \sim \mu^\star$. Denote the joint state and action as $s_t^N := (s_t^1, \dots, s_t^n) \in \mathcal{S}^N$ and $a_t^N := (a_t^1, \dots, a_t^n) \in \mathcal{A}^N$. At each time-step t, the empirical state distribution $e^N \in \Delta(\mathcal{S})$ is

$$e^N(s_t^N) := \frac{1}{n} \sum_{i=1}^{n} \delta_{s_t^N = s_t^i}, \quad \forall s_t^N \in \mathcal{S}. \tag{6}$$

For a pair $(s^N, a^N) \in \mathcal{S}^N \times \mathcal{A}^N$, the environment transition will follow some kernel from the uncertainty set under the empirical state distribution $e^N(\cdot)$

$$\mathfrak{P}^N(s^N, a^N) := \bigotimes_{i=1}^{n} \mathfrak{P}(s^i, a^i, e^N(s^N)). \tag{7}$$

In the game, let $\Pi := \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ and $\Pi^N := \Pi^N$ be profiles $\pi^N = (\pi^1, \dots, \pi^N)$ with $\pi^i \in \Pi$. For s^N, the joint action distribution from the joint (product) policy π^N is

$$\pi^N(da^N \mid s^N) := \bigotimes_{i=1}^{n} \pi^i(da^i \mid s^i). \tag{8}$$

Given a profile $\pi^N \in \Pi^N$, consider sequences of (possibly time-inhomogeneous) transition kernels $(p_t^N)_{t \geq 0}$ on $\mathcal{S}^N$ such that for every t and every $(s^N, a^N) \in \mathcal{S}^N \times \mathcal{A}^N$,

$$p_t^N(\cdot \mid s^N, a^N) \ \in \ \mathfrak{P}^N(s^N, a^N). \tag{9}$$

Given $(\pi^N, (p_t^N)_{t \geq 0})$, let $\mathbb{P}^{\pi^N, (p_t^N)}$ denote the induced law of the controlled process generated by $s_0^N \sim (\mu^\star)^{\otimes N}$, $a_t^N \sim \pi^N(\cdot \mid s_t^N)$, and $s_{t+1}^N \sim p_t^N(\cdot \mid s_t^N, a_t^N)$. We write $\mathbb{E}^{\pi^N, (p_t^N)}$ for expectation under $\mathbb{P}^{\pi^N, (p_t^N)}$.

The robust payoff of player i is then defined as

$$\begin{aligned} &J_i^N(\pi^N) \\ &:= \inf_{(p_t^N)} \ \mathbb{E}^{\pi^N, (p_t^N)} \Big[\sum_{t \geq 0} \gamma^t \, r(s_t^i, a_t^i, s_{t+1}^i, e^N(s_t^N)) \Big], \end{aligned} \tag{10}$$

where the infimum ranges over all sequences $(p_t^N)_{t\geq 0}$ satisfying the constraint above.

With this objective, we can define the Nash equilibrium notions as in standard games.

Definition 3. *A profile $\pi^{N,\star} \in \Pi^N$ is an ε-Nash equilibrium if for all i, $J_i^N(\pi^{N,\star}) + \varepsilon \geq \sup_{\pi\in\Pi} J_i^N(\pi^{N,\star,-i}, \pi)$, where $(\pi^{N,\star,-i}, \pi)$ is the joint policy where agent i takes π while others follow $\pi^{N,\star}$.*

Remark 2. *The N-player robust games we defined are closely related to standard robust Markov games [Zhang et al., 2020, Kardeş et al., 2011] (see Section B.1).*

6.2 Asymptotic Approximation of N-Player Robust Games

In this section, we show that the robust MFG provides an asymptotic approximation of the finite-player robust games defined above. Specifically, we show that, when the number of agents is sufficiently large, the equilibrium policy of the robust MFG is an ε-Nash equilibrium of the N-player robust game.

Define the symmetric equilibrium profile

$$\pi^{N|\star} := (\pi^\star, \dots, \pi^\star) \in \Pi^N. \tag{11}$$

Fix any unilateral deviation sequence $(\pi^{(N)})_{N\in\mathbb{N}} \subset \Pi$ for player 1, and set

$$\pi^{N|(N)} := (\pi^{(N)}, \pi^\star, \dots, \pi^\star) \in \Pi^N. \tag{12}$$

For each N, let $P^{N|(N)}$ be a minimizing law attaining $J_1^N(\pi^{N|(N)})$, whose existence is guaranteed by Lemma 8 in Appendix C. Under $P^{N|(N)}$, define the one-step law

$$Q_t^{N|(N)} := \mathrm{Law}_{P^{N|(N)}}\big(s_t^1, a_t^1, s_{t+1}^1, e^N(s_t^N)\big). \tag{13}$$

Next, define the proxy chain $P^{*(N)}$ by

$$s_0 \sim \mu^\star, a_t \sim \pi^{(N)}(\cdot \mid s_t), s_{t+1} \sim p^\star(\cdot \mid s_t, a_t), \tag{14}$$

and let

$$Q_t^{*(N)} := \mathrm{Law}_{P^{*(N)}}\big(s_t, a_t, s_{t+1}, \mu^\star\big), \qquad t \in \mathbb{N}_0. \tag{15}$$

Assumption 2. *For every deviation sequence $(\pi^{(N)})_N \subset \Pi$, every $t \in \mathbb{N}_0$, and some choice of minimizing laws $P^{N|(N)}$,*

$$W_1\big(Q_t^{N|(N)}, Q_t^{*(N)}\big) \longrightarrow 0 \qquad (N \to \infty), \tag{16}$$

with W_1 on $\Delta(Z)$ as in (6.3).

This assumption states that, along any unilateral deviation sequence, the deviating player asymptotically faces the same one-step law as in the proxy mean-field model driven by $(\mu^\star, \pi^{(N)}, p^\star)$. In particular, both the player-side transition and the population term entering the reward converge, at the level of one-step distributions, to their mean-field counterparts.

In the robust setting, such a stabilization property does not follow from Assumption 1 alone, because nature can use the remaining $N-1$ players to alter the empirical distribution in a way that persists at the level of the one-step law. The following counterexample shows that this additional assumption is genuinely needed.

Theorem 2. *Fix $\gamma \in (0,1)$ and $\kappa > 1/\gamma$, and set*

$$\varepsilon_0 \;:=\; \frac{\gamma\kappa - 1}{1-\gamma} \;>\; 0.$$

There exist a robust mean-field game satisfying Assumption 1 and a stationary robust mean-field equilibrium $(\mu^\star, \pi^\star, p^\star)$ of it such that:

- *(a) Assumption 2 fails for the constant deviation sequence $\pi^{(N)} \equiv \pi^\star$ under* every *choice of minimizing laws $P^{N|(N)}$; and*
- *(b) for* every *$N \in \mathbb{N}$, the symmetric profile $\pi^{N|\star}$ is not an ε-Nash equilibrium of the N-player robust game for any $\varepsilon < \varepsilon_0$.*

We then derive the asymptotic approximation result.

Theorem 3. *Assume Assumptions 1 and 2. Let $(\mu^\star, \pi^\star, p^\star)$ be a stationary robust mean-field equilibrium and let*

$$\pi^{N|\star} := (\pi^\star, \dots, \pi^\star) \in \Pi^N. \tag{17}$$

Then, for every $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that, for all $N \geq N(\varepsilon)$, the profile $\pi^{N|\star}$ is an ε-Nash equilibrium of the N-player robust game.

6.3 Non-Asymptotic Approximation

We now derive finite-N approximation bounds. We will mainly study the difference between the actual N-player robust game and the proxy chain driven by the equilibrium worst-case kernel $p^\star$. Specifically, let

$$X := \mathcal{S} \times \mathcal{A} \times \mathcal{S}, \qquad Z := X \times \Delta(\mathcal{S}), \tag{18}$$

and equip Z with the metric

$$d_Z\big((x,\nu),(\tilde{x},\tilde{\nu})\big) := \mathbf{1}_{\{x \neq \tilde{x}\}} + \|\nu - \tilde{\nu}\|_1, \tag{19}$$

and let W_1 be the corresponding Wasserstein-1 distance on $\mathcal{P}(Z)$.

In contrast with the asymptotic approximation result, the finite-N analysis requires an explicit rate rather than mere weak convergence. We therefore impose a quantitative hypothesis directly on the one-step law faced by the deviating player. Our assumption does not require a particular realization of the minimizing N-player kernel, and it only controls the law of the local transition together with the empirical distribution.

Assumption 3. *There exists a deterministic array $(\delta_{N,t})_{N\in\mathbb{N},\, t\in\mathbb{N}_0}$ with $\delta_{N,t} \geq 0$ such that, for every $N \in \mathbb{N}$, every unilateral deviation policy $\pi \in \Pi$, and every $t \in \mathbb{N}_0$,*

$$W_1\Big(Q_t^{N,\pi}, Q_t^{\pi}\Big) \leq \delta_{N,t}, \tag{20}$$

where $Q_t^{N,\pi}$ is the law of $\big(s_t^1, a_t^1, s_{t+1}^1, e^N(s_t^N)\big)$ under the actual N-player robust game with profile $\pi^{N|\star,-1} \oplus \pi$, and Q_t^{π} is the law of $\big(s_t, a_t, s_{t+1}, \mu^\star\big)$ under the proxy chain

$$s_0 \sim \mu^\star, \quad a_t \sim \pi(\cdot \mid s_t), \quad s_{t+1} \sim p^\star(\cdot \mid s_t, a_t). \tag{21}$$

Remark 3. *Assumption 3 is the quantitative counterpart of the one-step law convergence used in the asymptotic approximation. It is formulated directly at the level of*

$$Z_t^{N,\pi} := (s_t^1, a_t^1, s_{t+1}^1, \mu_t^N)$$

under the deviating N-player robust game and the proxy variable

$$Z_t^{\pi} := (s_t, a_t, s_{t+1}, \mu^\star)$$

under the mean-field worst-case kernel $p^\star$.

We note that the projections $(x,\nu) \mapsto \nu$ and $(x,\nu) \mapsto x$ are 1-Lipschitz under the product metric used in the assumption. Consequently, Assumption 3 simultaneously controls the empirical distribution and the local one-step dynamics. In particular,

$$W_1\big(\mathrm{Law}(\mu_t^N), \delta_{\mu^\star}\big) \leq \delta_{N,t}, \text{ hence } \mathbb{E}\big[\|\mu_t^N - \mu^\star\|_1\big] \leq \delta_{N,t},$$

and also

$$W_1\Big(\mathrm{Law}(s_t^1, a_t^1, s_{t+1}^1), \mathrm{Law}(s_t, a_t, s_{t+1})\Big) \leq \delta_{N,t},$$

where on $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ we use the discrete metric. Since the latter space is finite, this is equivalent to total-variation control of the local triple. Thus $\delta_{N,t}$ measures a combined finite-N Nash-certainty-equivalence error for both the deviator and the population.

Finally, the bound is required uniformly over unilateral deviations π, because the Nash gap involves the supremum over all such deviations. The dependence of $\delta_{N,t}$ on t is also natural: finite-N coupling errors may accumulate with time, and the theorem only needs the discounted series in the final bound to remain finite.

We further make the following standard Lipschitz assumption on reward, and derive the following finite-N equilibrium guarantee[2].

Assumption 4. *There exist $L_r > 0$ such that, for all $s \in S$, $a \in A$, $s' \in S$, and all $\nu, \tilde{\nu} \in \Delta(\mathcal{S})$,*

$$|r(s,a,s',\nu) - r(s,a,s',\tilde{\nu})| \leq L_r \|\nu - \tilde{\nu}\|_1. \tag{22}$$

Theorem 4. *Assume Assumptions 1, 3, and 4. Let $(\mu^\star, \pi^\star, p^\star)$ be a stationary robust mean-field equilibrium and let*

$$\pi^{N|\star} := (\pi^\star, \ldots, \pi^\star) \in \Pi^N. \tag{23}$$

[2]In Section C, we derive the results under a more general Hölder assumption.

Then $\pi^{N|\star}$ is an ε_N-Nash equilibrium of the N-player robust game, where

$$\varepsilon_N := 2\sum_{t=0}^{\infty}\gamma^t\Big(2\|r\|_\infty + L_r\Big)\delta_{N,t}. \tag{24}$$

Moreover, if there exists $C_\delta > 0$ such that[3]

$$\delta_{N,t} \le \frac{C_\delta(1+t)}{\sqrt{N}} \qquad \forall\, N \in \mathbb{N},\ t \in \mathbb{N}_0, \tag{25}$$

then

$$\varepsilon_N \le \frac{2(2\|r\|_\infty + L_r)C_\delta}{(1-\gamma)^2\sqrt{N}} = \mathcal{O}\Big(N^{-1/2}\Big). \tag{26}$$

7 Algorithmic Solutions for Robust MFGs

In this section, we further propose an iterative algorithm to compute a mean-field equilibrium of the robust MFG. We will first develop a fixed-point characterization of the mean-field equilibrium, and then design an algorithm with convergence guarantees.

7.1 Mean-Field Equilibria as Fixed Points

Fix $\mu \in \Delta(\mathcal{S})$. For $v \in \mathbb{R}^S$ define the robust Q-operator as

$$(Q_\mu v)(s,a) \triangleq \min_{P\in\mathfrak{P}(s,a,\mu)} \sum_{s'\in S} P(s')\Big(r(s,a,s',\mu) + \gamma v(s')\Big),$$

and the robust Bellman operator as

$$(T_\mu v)(s) \triangleq \max_{a\in A}(Q_\mu v)(s,a). \tag{27}$$

Since T_μ is a γ-contraction in $\|\cdot\|_\infty$, it admits a unique fixed point v_μ satisfying $v_\mu = T_\mu v_\mu$, and we further define the optimal robust Q-function by $Q_\mu(s,a) := (Q_\mu v_\mu)(s,a)$.

Moreover, a robust best response at μ can be obtained by selecting

$$a_\mu(s) \in \arg\max_{a\in A} Q_\mu(s,a), \qquad \pi_\mu(\cdot \mid s) = \delta_{a_\mu(s)}, \tag{28}$$

and for each (s,a), set $p_\mu(\cdot \mid s,a)$ as

$$\arg\min_{P\in\mathfrak{P}(s,a,\mu)} \mathbb{E}_{s'\sim P}\Big[r(s,a,s',\mu) + \gamma v_\mu(s')\Big]. \tag{29}$$

This triple (μ, π_μ, p_μ) hence induces a Markov kernel on $\mathcal{S}$:

$$K_\mu(s' \mid s) := \sum_{a\in A} \pi_\mu(a \mid s)\, p_\mu(s' \mid s,a), \tag{30}$$

and we define an operator $F : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ as

$$F(\mu) := \mu K_\mu. \tag{31}$$

We then prove that, any fixed point of F is a MFE.

Lemma 1. *Let $\mu \in \Delta(\mathcal{S})$ and suppose (π_μ, p_μ) are selected as in* (28)–(29)*. If μ satisfies the fixed-point condition*

$$\mu = F(\mu) = \mu K_\mu, \tag{32}$$

then (μ, π_μ, p_μ) is a stationary robust MFE.

This result thus enables us to reduce the problem of solving a robust MFG to finding a fixed point of F. Based on this, we design our algorithm as follows.

[3]The discussion of this condition is deferred to Section C.5.

Algorithm 1 Robust Best-Response Iteration

Input initial distribution $\mu^0 \in \Delta(\mathcal{S})$; stepsize α; DP tolerance $\varepsilon_{\mathrm{DP}} > 0$; mean-field tolerance $\varepsilon_\mu > 0$.
for $k = 0, 1, 2, \dots$ **do**
  Initialize $v^{(0)} \in \mathbb{R}^S$, $m \leftarrow 0$
  while $\|v^{(m+1)} - v^{(m)}\|_\infty > \varepsilon_{\mathrm{DP}}$ **do**
    $v^{(m+1)} \leftarrow T_{\mu^k} v^{(m)}$
    $m \leftarrow m + 1$
  end while
  $v_{\mu^k} \leftarrow v^{(m)}$
  Select $a^k(s) \in \arg\max_{a \in A} (Q_{\mu^k} v_{\mu^k})(s, a)$
  $\pi^k(\cdot \mid s) \leftarrow \delta_{a^k(s)}$
  Select $p^k(\cdot \mid s, a) \in \arg\min_{P \in \mathfrak{P}(s,a,\mu^k)} \sum_{s' \in S} P(s') \Big(r(s, a, s', \mu^k) + \gamma v_{\mu^k}(s') \Big)$.
  $K^k(s' \mid s) \leftarrow \sum_{a \in A} \pi^k(a \mid s) p^k(s' \mid s, a)$
  $\tilde{\mu}^{k+1} \leftarrow \mu^k K^k$.
  $\mu^{k+1} \leftarrow (1 - \alpha)\mu^k + \alpha \tilde{\mu}^{k+1}$.
  if $\|\mu^{k+1} - \mu^k\|_1 \le \varepsilon_\mu$ **then**
    break
  end if
end for
return (μ^{k+1}, π^k, p^k).

7.2 Convergence Analysis

Then we derive the convergence analysis of the algorithm. We first highlight that, even in a non-robust stationary MFG, it is not always guaranteed that F has a fixed point [Huang et al., 2006], and additional structural assumptions are needed, e.g., [Cardaliaguet and Hadikhanloo, 2017, Perrin et al., 2020, Geist et al., 2021, Perolat et al., 2022]. We hence adapt the standard contraction assumption as [4].

Assumption 5. *For a Markov kernel K, define its Dobrushin coefficient by*

$$\alpha(K) \;:=\; \max_{s, \tilde{s} \in S} d_{\mathrm{TV}}\big(K(\cdot \mid s), K(\cdot \mid \tilde{s})\big). \tag{33}$$

Assume there exist constants $\rho_{\mathrm{mix}} \in [0, 1)$ and $L_K \ge 0$ such that for all $\mu, \tilde{\mu} \in \Delta(\mathcal{S})$:

$$\alpha(K_\mu) \le \rho_{\mathrm{mix}}, \tag{34}$$

$$\max_{s \in S} \|K_\mu(\cdot \mid s) - K_{\tilde{\mu}}(\cdot \mid s)\|_1 \le L_K \|\mu - \tilde{\mu}\|_1. \tag{35}$$

And it holds that $\rho \;:=\; \rho_{\mathrm{mix}} + L_K \;<\; 1$.

Notably, as discussed in Section D.1, a broad range of distributional uncertainty sets satisfy the approximate Lipschitz condition, and thus our results hold for these sets with small Lipschitz coefficients.

Remark 4. *The assumption is standard and inevitable even in non-robust MFGs [Huang et al., 2006, Carlini and Silva, 2014, Guo et al., 2019, Anahtarci et al., 2023, Yardim et al., 2024, Yang, 2026], and any best-response-based or iteration-typed algorithms can be unstable without it [Cui and Koeppl, 2021, Chassagneux et al., 2019, Lauriere, 2021]. We hence adapt this assumption to our robust setting.*

On the other hand, there are extensive studies developing different techniques to ensure stability and convergence under weaker conditions, e.g., [Cardaliaguet and Hadikhanloo, 2017, Hadikhanloo and Silva, 2019, Hadikhanloo, 2018, Perrin et al., 2020, Geist et al., 2021, Lavigne and Pfeiffer, 2023, Delarue and Vasileiadis, 2025, Subramanian and Mahajan, 2019, Mguni et al., 2018, Angiuli et al., 2022, Xie et al., 2021]. We leave adaptations of these techniques to robust MFGs as our future work.

Under this assumption, we can prove the convergence of our Algorithm 1. Note that the algorithm updates may not be exact due to tolerance errors. We thus develop our analysis with an implemented population update $\widehat{F}(\mu^k)$ satisfying

$$\|\widehat{F}(\mu^k) - F(\mu^k)\|_1 \le \varepsilon_k. \tag{36}$$

[4] To make Assumption 5 more applicable, we further extend our algorithm design and results by replacing $\arg\max$ in (28) by a softmax policy, which allows us to ensure convergence under weaker and more verifiable assumptions. See our discussion in Section D.4.

Theorem 5. *Assume Assumption 5 and that Algorithm 1 uses the inexact update* $\mu^{k+1} = (1-\alpha)\mu^k + \alpha\widehat{F}(\mu^k)$ *with* (36)*. Then, denoting* $q := 1-\alpha(1-\rho) < 1$*, it holds that*

$$\|\mu^k - \mu^\star\|_1 \;\le\; q^k\|\mu^0 - \mu^\star\|_1 \;+\; \alpha\sum_{j=0}^{k-1} q^{k-1-j}\varepsilon_j. \tag{37}$$

In particular, if $\sup_j \varepsilon_j \le \bar{\varepsilon}$*, then* $\limsup_{k\to\infty}\|\mu^k - \mu^\star\|_1 \le \bar{\varepsilon}/(1-\rho)$.

Our result hence implies that Algorithm 1 will converge to the population distribution $\mu^\star$ of a mean-field equilibrium. Moreover, by learning the corresponding optimal robust policy and the worst-case kernel $(\pi_{\mu^\star}, p_{\mu^\star})$ under $\mu^\star$, we obtain a mean-field equilibrium. Our algorithm thus stands as the first concrete algorithm for stationary infinite-horizon robust MFGs with convergence guarantees.

Remark 5. *In our algorithm, computing a robust MFE thus reduces to iterating between (1) solving a robust MDP at fixed* μ *via contraction-based methods, and (2) updating* μ *via the induced Markov kernel. For a robust MDP, each robust Bellman update cost is polynomial in* S, A *[Iyengar, 2005, Nilim and El Ghaoui, 2004]. Thus, the total complexity of solving a robust MFG does not scale with* N*, whereas the complexity of solving a standard robust Markov game generally scales as* A^N *[Farhat et al., 2026]. Thus, our robust MFG provides a formulation of large MAS under model mismatches and circumvents the curse of multi-agency.*

8 Numerical Experiments

In this section, we develop numerical experiments to validate our theoretical results.

Experiment environments. In our experiments, we construct a robust MFG with $\mathcal{S} = \{0, 1, \ldots, S-1\}$ and $\mathcal{A} = \{0, 1\}$. Action $a = 0$ corresponds to *stay*, while $a = 1$ corresponds to *move clockwise* (to $s+1$) with slip.

Let $u \in \Delta(\mathcal{S})$ denote the uniform distribution on $\mathcal{S}$ and let $\eta \in (0, 1)$ be a mixing parameter. We first define transition kernels $\bar{p}_0(\cdot \mid s, a)$ under different (s, a) pairs, and set the nominal transition kernel as the uniformly-mixed kernel $p_0(\cdot \mid s, a) \;:=\; (1-\eta)\,\bar{p}_0(\cdot \mid s, a) \;+\; \eta\, u(\cdot)$.

Fix a goal state $s_g \in \mathcal{S}$ and parameters $R_g > 0$ (goal reward), $\lambda_{\text{cong}} \ge 0$ (congestion strength),and action costs $c(0), c(1) \ge 0$. We set the one-step reward as $r(s, a, s', \mu) \;:=\; R_g\,\mathbf{1}\{s' = s_g\} \;-\; \lambda_{\text{cong}}\,\mu(s') \;-\; c(a)$.

We evaluate robustness under two rectangular ambiguity models defined by D being L_1-norm and KL-divergence (for simplicity, we set them to be independent of μ): $\mathfrak{P}_(s, a) := \Big\{p :\; D(p, p_0(\cdot \mid s, a)) \le \rho\Big\}$.

Experiment results. We then develop two experimental protocols to validate our theoretical results.

Robustness of robust MFGs. Under both uncertainty sets, we will first obtain the robust MFE through our Algorithm 1 under a specific radius, and evaluate the robust value function of it under robust MFGs with different radii. Moreover, we find the non-robust MFE and plot its robust value under these robust MFGs as baselines.

We plot the robust value v.s. the uncertainty set radius in Figure 1. As shown by the results, our robust MFE is much more reliable and stable when facing model uncertainties, whereas the vanilla MFE suffers from severe performance degradations under model mismatches. These experiments hence verify the enhanced robustness and stability of our robust MFG formulation, validating our theoretical results.

Approximation of finite-player robust games. We then fix an ambiguity radius and compute the corresponding robust MFE $(\mu^\star, \pi^\star, p^\star)$. Then, for each N, we compute the Nash gap of $\pi^\star$ under $\mu^\star$ and the corresponding N-player robust game, and plot the gap v.s. the number of players.

Our results are presented in Figure 2. As shown in the results, when the player number N increases, the Nash gap of the robust MFE diminishes, i.e., it becomes an ε-Nash equilibrium in the N-player game. Moreover, the approximation slope is $\frac{1}{2}$ in the $\log$-scale, validating our non-asymptotic convergence rate. These results hence validate our theoretical results.

9 Conclusion

In this paper, we proposed and investigated stationary robust mean-field games (MFGs). We first proved its solvability, and developed comprehensive studies on the approximation of the robust MFGs to finite-agent robust games, revealing both asymptotic and non-asymptotic approximations as the multi-agent system becomes large. We further designed a

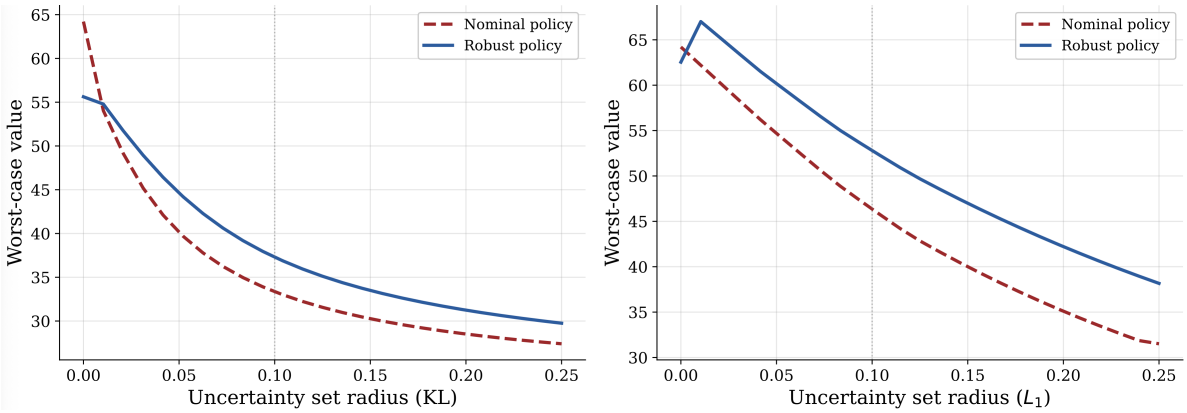

Figure 1: Robust MFE v.s. Non-robust MFE

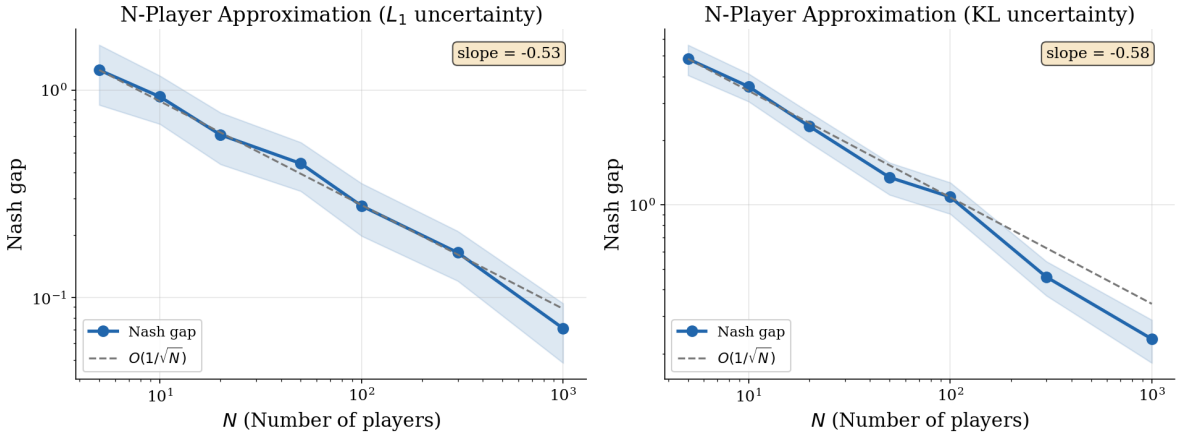

Figure 2: Approximation of N-Player Robust Games

concrete algorithm for robust MFGs with provable convergence guarantees. Our studies provide a tractable foundation for large-population multi-agent systems under model mismatches, which serve as a potential efficient solution to close the Sim-to-Real gap in large MAS.

Acknowledgements

This work was supported by an Amazon Research Award, Fall 2025. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Amazon.

References

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.

Sachin Adlakha, Ramesh Johari, and Gabriel Y. Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *J. Econom. Theory*, 156:269–316, 2015.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

Alexander Aurell, René Carmona, Gokce Dayanikli, and Mathieu Laurière. Optimal incentives to mitigate epidemics: a Stackelberg mean field game approach. *SIAM J. Control Optim.*, 60(2):S294–S322, 2022.

Dario Bauso, Hamidou Tembine, and Tamer Başar. Robust mean field games. *Dynamic games and applications*, 6(3): 277–303, 2016.

Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean field games and mean field type control theory*, volume 101. New York: Springer-Verlag, 2013.

Anup Biswas. Mean field games with ergodic cost for discrete time Markov processes. *arXiv preprint arXiv:1510.08968*, 2015.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36:66845–66859, 2023.

Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11): 4948, 2021.

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.

Elisabetta Carlini and Francisco J Silva. A fully discrete semi-lagrangian scheme for a first order mean field game problem. *SIAM Journal on Numerical Analysis*, 52(1):45–67, 2014.

René Carmona and François Delarue. Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, 51(4):2705–2734, 2013.

René Carmona, François Delarue, et al. *Probabilistic theory of mean field games with applications I-II*, volume 3. Springer, 2018.

Jean-François Chassagneux, Dan Crisan, and François Delarue. Numerical method for fbsdes of mckean–vlasov type. *The Annals of Applied Probability*, 29(3):1640–1684, 2019.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. *Advances in Neural Information Processing Systems*, 38:85402–85463, 2026.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

François Delarue and Athanasios Vasileiadis. Exploration noise for learning linear-quadratic mean field games. *Mathematics of Operations Research*, 50(3):1762–1831, 2025.

François Delarue, Daniel Lacker, and Kavita Ramanan. From the master equation to mean field game limit theory. *The Annals of Probability*, 48(1):211–263, 2020.

Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7143–7150, 2020.

Robert Elliott, Xun Li, and Yuan-Hua Ni. Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, 49(11):3222–3233, 2013.

Zain Ulabedeen Farhat, Debamita Ghosh, George K Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In *The Fourteenth International Conference on Learning Representations*, 2026.

Nicolas Gast, Bruno Gaujal, and Jean-Yves Le Boudec. Mean field for Markov decision processes: from discrete to continuous optimization. *IEEE. Trans. Autom. Control*, 57(9):2266–2280, 2012.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

Debamita Ghosh, George K Atia, and Yue Wang. Scaling online distributionally robust reinforcement learning: Sample-efficient guarantees with general function approximation. *arXiv preprint arXiv:2512.18957*, 2025.

Debamita Ghosh, George K Atia, and Yue Wang. Orvit: Near-optimal online distributionally robust reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21278–21286, 2026.

Diogo A Gomes and João Saúde. Mean field games models—A brief survey. *Dynamic Games and Applications*, 4(2): 110–154, 2014.

Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, 93(3):308–328, 2010.

Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. Continuous time finite state mean field games. *Appl. Math. Optim.*, 68(1):99–143, 2013.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.

Saeed Hadikhanloo. *Learning in mean field games*. PhD thesis, Université Paris sciences et lettres, 2018.

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.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

Min Hua, Xinda Qi, Dong Chen, Kun Jiang, Zemin Eitan Liu, Hongyu Sun, Quan Zhou, and Hongming Xu. Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects. *IEEE Transactions on Automation Science and Engineering*, 2025.

Jianhui Huang and Minyi Huang. Mean field LQG games with model uncertainty. In *52nd IEEE Conference on Decision and Control*, pages 3103–3108. IEEE, 2013.

Minyi Huang. Large-population lqg games involving a major player: the nash certainty equivalence principle. *SIAM Journal on Control and Optimization*, 48(5):3318–3353, 2010.

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.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Yuchen Jiao and Gen Li. Minimax-optimal multi-agent robust reinforcement learning. *arXiv preprint arXiv:2412.19873*, 2024.

Erim Kardeş, Fernando Ordóñez, and Randolph W Hall. Discounted robust stochastic games and an application to queueing control. *Operations research*, 59(2):365–382, 2011.

Daniel Lacker. A general characterization of the mean field limit for stochastic differential games. *Probab. Theory Relat. Fields*, 165:581–648, 2016.

Daniel Lacker and Agathe Soret. A case study on stochastic games on large graphs in mean field and sparse regimes. *Math. Oper. Res.*, 47(2):1530–1565, 2022.

Daniel Lacker and Thaleia Zariphopoulou. Mean field and n-agent games for optimal investment under relative performance criteria. *Math. Finance*, 29(4):1003–1038, 2019.

Johannes Langner, Ariel Neufeld, and Kyunghyun Park. Markov-nash equilibria in mean-field games under model uncertainty. *arXiv preprint arXiv:2410.11652*, 2024.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japan. J. Math.*, 2(1):229–260, 2007.

Mathieu Lauriere. Numerical methods for mean field games and mean field type control. *arXiv preprint arXiv:2106.06231*, 2021.

Mathieu Laurière, Sarah Perrin, Julien Pérolat, Sertan Girgin, Paul Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Learning in mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022.

Mathieu Laurière, Ariel Neufeld, and Kyunghyun Park. Robust mean-field control under common noise uncertainty. *arXiv preprint arXiv:2511.04515*, 2025.

Pierre Lavigne and Laurent Pfeiffer. Generalized conditional gradient and learning in potential mean field games. *Applied Mathematics & Optimization*, 88(3):89, 2023.

Na Li, Zewu Zheng, Wei Ni, Hangguan Shan, Wenjie Zhang, and Xinyu Li. Sample-efficient tabular self-play for offline robust reinforcement learning. *Advances in Neural Information Processing Systems*, 38:155562–155626, 2026.

Zongxia Liang, Zhou Zhou, Yaqi Zhuang, and Bin Zou. Mean-field games under model uncertainty. *arXiv preprint arXiv:2601.12226*, 2026.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Guangyi Liu, Suzan Iloglu, Michael Caldara, Joseph W Durham, and Michael M. Zavlanos. Distributionally robust multi-agent reinforcement learning for dynamic chute mapping. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 13623–13643. PMLR, 2022.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 6379–6390, 2017.

Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. *Advances in Neural Information Processing Systems*, 37:12528–12580, 2024.

Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

David Mguni, Joel Jennings, and Enrique Munoz de Cote. Decentralised learning in systems with many, many strategic agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jun Moon and Tamer Başar. Discrete-time decentralized control using the risk-sensitive performance criterion in the large population regime: a mean field approach. In *ACC 2015*. Chicago, 2015.

Jun Moon and Tamer Başar. Discrete-time stochastic stackelberg dynamic games with a large number of followers. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3578–3583. IEEE, 2016a.

Jun Moon and Tamer Başar. Linear quadratic risk-sensitive and robust mean field games. *IEEE. Trans. Autom. Control*, 62(3):1062–1077, 2016b.

Jun Moon and Tamer Başar. Robust mean field games for coupled Markov jump linear systems. *Internat. J. Control*, 89(7):1367–1381, 2016c.

Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 839–846, 2004.

Mojtaba Nourian and Girish N Nair. Linear-quadratic-gaussian mean field games under high rate quantization. In *52nd IEEE Conference on Decision and Control*, pages 1898–1903. IEEE, 2013.

Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9582–9602. PMLR, 2022.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems*, 33:13199–13213, 2020.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 2817–2826. PMLR, 2017.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016.

Zachary Roch, George Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2025a.

Zachary Roch, Chi Zhang, George Atia, and Yue Wang. A finite-sample analysis of distributionally robust average-reward reinforcement learning. *arXiv preprint arXiv:2505.12462*, 2025b.

Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate markov-nash equilibria for discrete-time risk-sensitive mean-field games. *Mathematics of Operations Research*, 45(4):1596–1620, 2020.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36:79903–79917, 2023.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proc. International Conference on Machine Learning (ICML)*, pages 44909–44959. PMLR, 2024.

Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 54904–54918. PMLR, 2025.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550 (7676):354–359, 2017.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, pages 251–259, 2019.

Hamidou Tembine, Quanyan Zhu, and Tamer Başar. Risk-sensitive mean-field games. *IEEE. Trans. Autom. Control*, 59(4):835–850, 2013.

Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. In *Reinforcement Learning Conference*, 2024a. URL `https://openreview.net/forum?id=OUiOUeVnb5`.

Shengbo Wang and Nian Si. Bellman optimality of average-reward robust markov decision processes with a constant gain. *arXiv preprint arXiv:2509.14203*, 2025.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3370–3398. PMLR, 2023a.

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024b.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 7193–7206, 2021.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward markov decision processes. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 37, pages 15215–15223, 2023b.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 36431–36469. PMLR, 2023c.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. A unified principle of pessimism for offline reinforcement learning under model mismatch. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward reinforcement learning. *Journal of Artificial Intelligence Research*, 80:719–803, 2024d.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6):5023–5056, 2023.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9728–9754. PMLR, 2023.

Zhenhui Xu, Jiayu Chen, Bing-Chang Wang, Yuhu Wu, and Tielong Shen. Robust mean field social control: a unified reinforcement learning framework. *arXiv preprint arXiv:2502.20029*, 2025.

Shan Yang. Mean-field reinforcement learning without synchrony, 2026. URL `https://arxiv.org/abs/2602.18026`.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR, 2023.

Batuhan Yardim, Artur Goldman, and Niao He. When is mean-field reinforcement learning tractable and relevant? In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2038–2046, 2024.

Muhammad Aneeq Uz Zaman, Mathieu Lauriere, Alec Koppel, and Tamer Başar. Robust cooperative multi-agent reinforcement learning: A mean-field type game perspective. In *6th Annual Learning for Dynamics & Control Conference*, pages 770–783. PMLR, 2024.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

A Stationary discounted robust mean-field games

A.1 Robust discounted dynamic programming

For fixed $\mu \in \Delta(\mathcal{S})$ and $v \in \mathbb{R}^{\mathcal{S}}$ define the robust Q-operator

$$(\mathcal{Q}_\mu v)(s,a) := \min_{P \in \mathfrak{P}(s,a,\mu)} \sum_{s' \in \mathcal{S}} P(s') \Big(r(s,a,s',\mu) + \gamma v(s') \Big). \tag{38}$$

Define the robust Bellman operator $\mathcal{T}_\mu : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ by

$$(\mathcal{T}_\mu v)(s) := \max_{a \in \mathcal{A}} (\mathcal{Q}_\mu v)(s,a). \tag{39}$$

Lemma 2. *For each fixed $\mu \in \Delta(\mathcal{S})$, the operator $\mathcal{T}_\mu$ is a contraction on $(\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$ with modulus γ. Hence there is a unique $v_\mu \in \mathbb{R}^{\mathcal{S}}$ such that*

$$v_\mu = \mathcal{T}_\mu v_\mu. \tag{40}$$

Proof. Fix $\mu \in \Delta(\mathcal{S})$ and $v, w \in \mathbb{R}^{\mathcal{S}}$. Fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For any $P \in \mathfrak{P}(s,a,\mu)$,

$$\begin{aligned} &\Big| \sum_{s'} P(s') \big(r(s,a,s',\mu) + \gamma v(s') \big) - \sum_{s'} P(s') \big(r(s,a,s',\mu) + \gamma w(s') \big) \Big| \\ &\quad = \gamma \Big| \sum_{s'} P(s') \big(v(s') - w(s') \big) \Big| \le \gamma \|v - w\|_\infty, \end{aligned}$$

since $\sum_{s'} P(s') = 1$ and $|v(s') - w(s')| \le \|v - w\|_\infty$ for all s'. Taking the minimum over $P \in \mathfrak{P}(s,a,\mu)$ gives $|(\mathcal{Q}_\mu v)(s,a) - (\mathcal{Q}_\mu w)(s,a)| \le \gamma \|v - w\|_\infty$. Taking the maximum over $a \in \mathcal{A}$ yields $|(\mathcal{T}_\mu v)(s) - (\mathcal{T}_\mu w)(s)| \le \gamma \|v - w\|_\infty$ for all s. Thus $\|\mathcal{T}_\mu v - \mathcal{T}_\mu w\|_\infty \le \gamma \|v - w\|_\infty$. Since $\gamma \in (0,1)$, Banach's fixed point theorem implies existence and uniqueness of v_μ solving (40). □

Define the optimal robust Q-function

$$Q_\mu(s,a) := (\mathcal{Q}_\mu v_\mu)(s,a), \tag{41}$$

so $v_\mu(s) = \max_{a \in \mathcal{A}} Q_\mu(s,a)$.

Lemma 3. *Fix $\mu \in \Delta(\mathcal{S})$.*

(i) For every $(s,a) \in \mathcal{S} \times \mathcal{A}$ the set

$$\widehat{\mathfrak{P}}(s,a,\mu) := \operatorname*{arg\,min}_{P \in \mathfrak{P}(s,a,\mu)} \sum_{s' \in \mathcal{S}} P(s') \big(r(s,a,s',\mu) + \gamma v_\mu(s') \big) \tag{42}$$

is nonempty, convex, and compact.

(ii) For every $s \in \mathcal{S}$ the set of optimal actions

$$\mathsf{D}(s,\mu) := \operatorname*{arg\,max}_{a \in \mathcal{A}} Q_\mu(s,a) \tag{43}$$

is nonempty.

Proof. (i) Fix (s,a). The objective is continuous and affine in P. By Assumption 1(ii), $\mathfrak{P}(s,a,\mu)$ is nonempty and compact, hence the minimum is attained and the argmin set is nonempty and compact. Convexity follows because $\mathfrak{P}(s,a,\mu)$ is convex and the objective is affine.

(ii) Since $\mathcal{A}$ is finite, the maximum of $a \mapsto Q_\mu(s,a)$ is attained, so $\mathsf{D}(s,\mu) \neq \emptyset$. □

Proposition 1. *Fix $\mu \in \Delta(\mathcal{S})$. Let v_μ be the unique fixed point of $\mathcal{T}_\mu$. Let π be any stationary policy satisfying $\operatorname{supp} \pi(\cdot|s) \subseteq \mathsf{D}(s,\mu)$ for all s, and let p be any stationary kernel with $p(\cdot|s,a) \in \widehat{\mathfrak{P}}(s,a,\mu)$ for all (s,a). Then:*

(i) For every $s \in \mathcal{S}$,

$$v_\mu(s) = \inf_{p' \in \mathfrak{P}(s,a,\mu)} J_\mu(s;\pi,p') = J_\mu(s;\pi,p). \tag{44}$$

(ii) For every stationary policy π',

$$\inf_{p'} J_\mu(s;\pi',p') \le v_\mu(s) \qquad \forall s \in \mathcal{S}. \tag{45}$$

In particular, v_μ coincides with the robust value (2) *and every π supported on the greedy sets attains the supremum in* (2)*.*

Proof. For any stationary pair (π,p) with $p(\cdot \mid s,a) \in \mathfrak{P}(s,a,\mu)$, define for $v \in \mathbb{R}^{\mathcal{S}}$:

$$(\mathcal{T}_\mu^{\pi,p} v)(s) := \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} \big(r(s,a,s',\mu) + \gamma v(s')\big) p(s' \mid s,a). \tag{46}$$

Exactly as in Lemma 2, $\mathcal{T}_\mu^{\pi,p}$ is a contraction with modulus γ, and it is affine. Hence it has a unique fixed point, call it $u_\mu^{\pi,p}$. Define partial sums

$$J_\mu^{(n)}(s;\pi,p) := \mathbb{E}_{s_0=s}^{\pi,p}\Big[\sum_{t=0}^{n-1} \gamma^t r(s_t,a_t,s_{t+1},\mu)\Big]. \tag{47}$$

Then $J_\mu^{(0)}(\cdot;\pi,p) \equiv 0$ and the Markov property yields the recursion

$$J_\mu^{(n+1)}(\cdot;\pi,p) = \mathcal{T}_\mu^{\pi,p} J_\mu^{(n)}(\cdot;\pi,p). \tag{48}$$

Thus $J_\mu^{(n)} = (\mathcal{T}_\mu^{\pi,p})^n 0 \to u_\mu^{\pi,p}$ in $\|\cdot\|_\infty$. Since r is bounded and $\gamma \in (0,1)$, $J_\mu^{(n)}(s;\pi,p) \to J_\mu(s;\pi,p)$ for each s. Therefore $u_\mu^{\pi,p} = J_\mu(\cdot;\pi,p)$.

Define

$$(\mathcal{T}_\mu^{\pi} v)(s) := \sum_{a\in\mathcal{A}} \pi(a \mid s) \min_{P\in\mathfrak{P}(s,a,\mu)} \sum_{s'\in\mathcal{S}} \big(r(s,a,s',\mu) + \gamma v(s')\big) P(s'). \tag{49}$$

Again, $\mathcal{T}_\mu^{\pi}$ is a contraction with modulus γ, hence has a unique fixed point u_μ^{π}.

We show $u_\mu^{\pi}(s) = \inf_p J_\mu(s;\pi,p)$. For any admissible p, by definition of the minimum, $\mathcal{T}_\mu^{\pi,p} v \ge \mathcal{T}_\mu^{\pi} v$ pointwise. Iterating from 0 gives $(\mathcal{T}_\mu^{\pi,p})^n 0 \ge (\mathcal{T}_\mu^{\pi})^n 0$ for all n[5]. Taking limits and using Step 1,

$$J_\mu(\cdot;\pi,p) = \lim_{n\to\infty} (\mathcal{T}_\mu^{\pi,p})^n 0 \ge \lim_{n\to\infty} (\mathcal{T}_\mu^{\pi})^n 0 = u_\mu^{\pi}. \tag{50}$$

Hence $\inf_p J_\mu(s;\pi,p) \ge u_\mu^{\pi}(s)$.

Conversely, for each (s,a) choose a minimizer $p^{\pi,u_\mu^\pi}(\cdot \mid s,a) \in \arg\min_{P\in\mathfrak{P}(s,a,\mu)} \sum_{s'} (r + \gamma u_\mu^\pi) P(s')$. Then $\mathcal{T}_\mu^{\pi} u_\mu^{\pi} = \mathcal{T}_\mu^{\pi,p^{\pi,u_\mu^\pi}} u_\mu^{\pi} = u_\mu^{\pi}$. By the result above, this implies $u_\mu^{\pi} = J_\mu(\cdot;\pi,p^{\pi,u_\mu^\pi})$ and hence $\inf_p J_\mu \le u_\mu^{\pi}$. Thus $\inf_p J_\mu = u_\mu^{\pi}$.

For every v, we have that

$$\mathcal{T}_\mu v = \max_{a\in\mathcal{A}} (\mathcal{Q}_\mu v)(\cdot,a) = \sup_\pi \mathcal{T}_\mu^{\pi} v, \tag{51}$$

where the supremum is attained by deterministic π choosing a maximizing action. Hence $\mathcal{T}_\mu^{\pi} v \le \mathcal{T}_\mu v$ for all π, so by contraction iteration from 0,

$$u_\mu^{\pi} \le v_\mu, \quad \forall \pi. \tag{52}$$

Thus $\sup_\pi u_\mu^{\pi}(s) \le v_\mu(s)$.

Now choose π supported on maximizers of Q_μ: $\operatorname{supp}\pi(\cdot|s) \subset \mathsf{D}(s,\mu)$ for any s. Then for each s,

$$\mathcal{T}_\mu^{\pi} v_\mu(s) = \sum_a \pi(a|s) \mathcal{Q}_\mu v_\mu(s,a) = \sum_a \pi(a|s) Q_\mu(s,a) = \max_a Q_\mu(s,a) = v_\mu(s), \tag{53}$$

where the third equality holds because $\pi(\cdot|s)$ puts all mass on $\arg\max_a Q_\mu(s,a)$. Thus v_μ is a fixed point of $\mathcal{T}_\mu^{\pi}$, hence by uniqueness $u_\mu^{\pi} = v_\mu$. Therefore $\sup_\pi u_\mu^{\pi}(s) \ge v_\mu(s)$ and $\sup_\pi u_\mu^{\pi}(s) = v_\mu(s)$.

Since $p(\cdot \mid s,a) \in \widehat{\mathfrak{P}}(s,a,\mu)$ for each (s,a). Then by construction,

$$(\mathcal{T}_\mu^{\pi,p} v_\mu)(s) = v_\mu(s) \quad \forall s, \tag{54}$$

so v_μ is the fixed point of $\mathcal{T}_\mu^{\pi,p}$ and it yields $J_\mu(\cdot;\pi,p) = v_\mu$. This proves (i). Part (ii) follows from $\sup_\pi \inf_p J_\mu = v_\mu$ proved above. □

[5]Here we use the monotonicity of the operators, proved in Lemma 4.

Lemma 4 (Monotonicity)**.** *Fix $\mu \in \Delta(\mathcal{S})$, a stationary policy π, and a stationary kernel p with $p(\cdot|s,a) \in \mathfrak{P}(s,a,\mu)$. Each of the operators $\mathcal{T}_\mu$, $\mathcal{T}_\mu^\pi$, $\mathcal{T}_\mu^{\pi,p}$ defined above is monotone: $v \le w$ componentwise implies $\mathcal{T}v \le \mathcal{T}w$. Moreover each commutes with constant shifts in the standard way: $\mathcal{T}(v + c\mathbf{1}) = \mathcal{T}v + \gamma c\,\mathbf{1}$ for $c \in \mathbb{R}$.*

Proof. For $\mathcal{T}_\mu^{\pi,p}$ both claims are immediate since the coefficients $\pi(a|s)p(s'|s,a) \ge 0$ sum (over s', for each a) to one. For the inner robust term, fix (s,a) and let $f_P(v) := \sum_{s'} P(s')\big[r(s,a,s',\mu) + \gamma v(s')\big]$. If $v \le w$ then $f_P(v) \le f_P(w)$ for every P. Let P_w attain $\min_P f_P(w)$. Then $\min_P f_P(v) \le f_{P_w}(v) \le f_{P_w}(w) = \min_P f_P(w)$, so $v \mapsto \min_{P \in \mathfrak{P}(s,a,\mu)} f_P(v)$ is monotone. Averaging over $\pi(\cdot|s)$ (for $\mathcal{T}_\mu^\pi$) or maximizing over a (for $\mathcal{T}_\mu$) preserves monotonicity. The shift property follows since $\sum_{s'} P(s')\gamma c = \gamma c$ for every $P \in \Delta(\mathcal{S})$, and shifts commute with $\min$, averages, and $\max$. □

Lemma 5. *Fix $\mu \in \Delta(\mathcal{S})$ and let (π^μ, p^μ) be as in Proposition 1. Define for $v \in \mathbb{R}^{\mathcal{S}}$ the (non-robust) Bellman operator with* fixed *kernel p^μ:*

$$(\widetilde{\mathcal{T}}_\mu v)(s) := \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p^\mu(s' \mid s,a)\big(r(s,a,s',\mu) + \gamma v(s')\big). \tag{55}$$

Then $\widetilde{\mathcal{T}}_\mu$ is a contraction with modulus γ and its unique fixed point equals v_μ. Consequently, for every initial distribution $\lambda \in \Delta(\mathcal{S})$,

$$\sum_{s \in \mathcal{S}} \lambda(s) v_\mu(s) = \sup_{\pi: \mathcal{S} \to \Delta(\mathcal{A})} \mathbb{E}^{\pi,p^\mu}_{s_0 \sim \lambda}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu)\Big], \tag{56}$$

and the supremum is attained at $\pi = \pi^\mu$.

Proof. **Step 1: $\widetilde{\mathcal{T}}_\mu$ is a contraction.** Fix $v, w \in \mathbb{R}^{\mathcal{S}}$. For any s and a,

$$\Big|\sum_{s'} p^\mu(s' \mid s,a)\gamma(v(s') - w(s'))\Big| \le \gamma \|v - w\|_\infty. \tag{57}$$

Taking the maximum over a gives $|(\widetilde{\mathcal{T}}_\mu v)(s) - (\widetilde{\mathcal{T}}_\mu w)(s)| \le \gamma\|v - w\|_\infty$. Thus $\|\widetilde{\mathcal{T}}_\mu v - \widetilde{\mathcal{T}}_\mu w\|_\infty \le \gamma\|v - w\|_\infty$.

Step 2: v_μ is a fixed point of $\widetilde{\mathcal{T}}_\mu$. By definition of $p^\mu(\cdot \mid s,a) \in \widehat{\mathfrak{P}}(s,a,\mu)$, for every (s,a),

$$(\mathcal{Q}_\mu v_\mu)(s,a) = \sum_{s'} p^\mu(s' \mid s,a)\big(r(s,a,s',\mu) + \gamma v_\mu(s')\big). \tag{58}$$

Therefore,

$$(\widetilde{\mathcal{T}}_\mu v_\mu)(s) = \max_a \sum_{s'} p^\mu(s' \mid s,a)\big(r + \gamma v_\mu\big) = \max_a (\mathcal{Q}_\mu v_\mu)(s,a) = (\mathcal{T}_\mu v_\mu)(s) = v_\mu(s). \tag{59}$$

So v_μ is a fixed point of $\widetilde{\mathcal{T}}_\mu$. By Step 1 and Banach's theorem, the fixed point is unique, hence it equals v_μ.

Step 3: Representation as an MDP value. Fix any stationary policy π. Define the policy-evaluation operator

$$(\widetilde{\mathcal{T}}_\mu^\pi v)(s) := \sum_a \pi(a \mid s) \sum_{s'} p^\mu(s' \mid s,a)\big(r(s,a,s',\mu) + \gamma v(s')\big). \tag{60}$$

It is a contraction with modulus γ and thus has a unique fixed point v^{π,p^μ}. The standard contraction iteration argument (identical to Step 1 of Proposition 1) shows that for each s,

$$v^{\pi,p^\mu}(s) = \mathbb{E}^{\pi,p^\mu}_{s_0=s}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu)\Big]. \tag{61}$$

By definition of $\widetilde{\mathcal{T}}_\mu$ as the pointwise maximum over actions, v_μ is the optimal value for the MDP with kernel p^μ, hence $v^{\pi,p^\mu}(s) \le v_\mu(s)$ for all s and equality holds for $\pi = \pi^\mu$. Averaging over $s_0 \sim \lambda$ gives (56). □

Note that under rectangularity, we will have that $V_\mu(\lambda) = \sum_s \lambda(s) v_\mu(s)$, as we proved in the following result.

Proposition 2. *Fix $\mu \in \Delta(\mathcal{S})$ and consider the rectangular admissible kernel class*

$$\mathcal{P}(\mu) := \Big\{ p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) : \; p(\cdot \mid s, a) \in \mathfrak{P}(s, a, \mu) \; \forall (s, a) \Big\}. \tag{62}$$

For $\lambda \in \Delta(\mathcal{S})$ define the distributional robust value

$$V_\mu(\lambda) := \sup_\pi \inf_{p \in \mathcal{P}(\mu)} \mathbb{E}^{\pi,p}_{s_0 \sim \lambda} \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu) \Big]. \tag{63}$$

Let $v_\mu : \mathcal{S} \to \mathbb{R}$ be the (unique) fixed point of the robust Bellman operator T_μ, equivalently $v_\mu(s) = \sup_\pi \inf_{p \in \mathcal{P}(\mu)} J_\mu(s; \pi, p)$ for each s. Then for every $\lambda \in \Delta(\mathcal{S})$,

$$V_\mu(\lambda) = \sum_{s \in \mathcal{S}} \lambda(s) \, v_\mu(s) =: \langle \lambda, v_\mu \rangle. \tag{64}$$

In particular, taking $\lambda = \mu$ gives $V_\mu(\mu) = \langle \mu, v_\mu \rangle$.

Proof. Fix $\lambda \in \Delta(\mathcal{S})$.

For any stationary policy π and admissible stationary kernel $p \in \mathcal{P}(\mu)$, by conditioning on s_0 we have

$$J_\mu(\lambda; \pi, p) = \mathbb{E}^{\pi,p}_{s_0 \sim \lambda} \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu) \Big] = \sum_{s \in \mathcal{S}} \lambda(s) \, J_\mu(s; \pi, p). \tag{65}$$

By Proposition 1 (robust discounted DP under (s, a)-rectangularity), there exist a stationary policy π^μ and an admissible stationary kernel $p^\mu \in \mathcal{P}(\mu)$ such that for every $s \in \mathcal{S}$,

$$v_\mu(s) = \inf_{p \in \mathcal{P}(\mu)} J_\mu(s; \pi^\mu, p) = J_\mu(s; \pi^\mu, p^\mu). \tag{66}$$

Multiplying (66) by $\lambda(s)$ and summing over s yields

$$\inf_{p \in \mathcal{P}(\mu)} J_\mu(\lambda; \pi^\mu, p) = J_\mu(\lambda; \pi^\mu, p^\mu) = \sum_s \lambda(s) v_\mu(s) = \langle \lambda, v_\mu \rangle. \tag{67}$$

Therefore, by definition of $V_\mu(\lambda)$,

$$V_\mu(\lambda) \geq \inf_{p \in \mathcal{P}(\mu)} J_\mu(\lambda; \pi^\mu, p) = \langle \lambda, v_\mu \rangle. \tag{68}$$

We then claim that for every stationary policy π and every state s,

$$J_\mu(s; \pi, p^\mu) \leq v_\mu(s). \tag{69}$$

Indeed, define the standard (non-robust) Bellman operator for the fixed kernel p^μ:

$$(\widetilde{T}_\mu v)(s) := \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p^\mu(s' \mid s, a) \big(r(s, a, s', \mu) + \gamma v(s') \big). \tag{70}$$

Because $p^\mu(\cdot \mid s, a)$ attains the inner minimum in the robust Bellman operator at v_μ, the robust fixed point identity $v_\mu = T_\mu v_\mu$ implies $v_\mu = \widetilde{T}_\mu v_\mu$. Since $\widetilde{T}_\mu$ is a contraction with modulus γ, v_μ is its unique fixed point, hence it is the optimal value function for the MDP with transition kernel p^μ. Therefore (69) holds.

Using (69) and Step 1, for any π we have

$$\inf_{p \in \mathcal{P}(\mu)} J_\mu(\lambda; \pi, p) \leq J_\mu(\lambda; \pi, p^\mu) = \sum_s \lambda(s) J_\mu(s; \pi, p^\mu) \leq \sum_s \lambda(s) v_\mu(s) = \langle \lambda, v_\mu \rangle. \tag{71}$$

Taking $\sup_\pi$ of the left-hand side yields $V_\mu(\lambda) \leq \langle \lambda, v_\mu \rangle$.

Thus it holds that $V_\mu(\lambda) \geq \langle \lambda, v_\mu \rangle$ and $V_\mu(\lambda) \leq \langle \lambda, v_\mu \rangle$, hence equality (64) holds. □

Proposition 3. *Fix $\mu \in \Delta(\mathcal{S})$ and a stationary policy π. We name $(q_t)_{t \geq 0}$ an admissible history-dependent adversary if each q_t maps histories $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ and actions a_t to $q_t(\cdot \,|\, h_t, a_t) \in \mathfrak{P}(s_t, a_t, \mu)$. Let $J_\mu(s; \pi, (q_t))$ be the induced discounted reward from $s_0 = s$. Then*

$$\inf_{(q_t) \text{ history-dep.}} J_\mu(s; \pi, (q_t)) = \inf_{p \text{ stationary}} J_\mu(s; \pi, p) = u^\pi_\mu(s) \qquad \forall s \in \mathcal{S}.$$

Proof. "$\leq$": every stationary kernel induces an admissible history-dependent adversary, so the left infimum is no larger; the right identity directly comes from the definition.

"$\geq$": fix any admissible (q_t) and write $u := u^{\pi}_{\mu}$. Conditioning on the history h_t (i.e., s_t is known, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim q_t(\cdot|h_t, a_t)$),

$$\mathbb{E}\big[r(s_t, a_t, s_{t+1}, \mu) + \gamma u(s_{t+1}) \,\big|\, h_t\big] = \sum_a \pi(a|s_t) \sum_{s'} q_t(s'|h_t, a)\big[r(s_t, a, s', \mu) + \gamma u(s')\big] \geq (\mathcal{T}^{\pi}_{\mu} u)(s_t) = u(s_t),$$

since $q_t(\cdot|h_t, a) \in \mathfrak{P}(s_t, a, \mu)$ and $\mathcal{T}^{\pi}_{\mu}$ takes the minimum over that set. Multiplying by γ^t, taking expectations, and summing for $t = 0, \ldots, T-1$ telescopes to $\mathbb{E}\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1}, \mu) \;\geq\; u(s) - \gamma^T \,\mathbb{E}\, u(s_T)$. Since $\|u\|_\infty \leq \|r\|_\infty/(1-\gamma)$, letting $T \to \infty$ then gives $J_\mu(s; \pi, (q_t)) \geq u(s)$. □

A.2 Existence of a stationary robust mean-field equilibrium

In this section we study existence of robust MFE.

Let

$$\Pi := \prod_{s \in S} \Delta(A), \qquad \mathcal{K} := \prod_{(s,a) \in S \times A} \Delta(S). \tag{72}$$

Thus an element $\pi \in \Pi$ is a stationary policy and an element $p \in \mathcal{K}$ is a stationary kernel slice $p(\cdot|s, a)$.

For fixed $\mu \in \Delta(S)$, let v_μ be the unique fixed point of T_μ, and define

$$Q_\mu(s, a) := \min_{P \in P(s,a,\mu)} \sum_{s' \in S} P(s')\big(r(s, a, s', \mu) + \gamma v_\mu(s')\big). \tag{73}$$

Set

$$D(s, \mu) := \arg\max_{a \in A} Q_\mu(s, a), \tag{74}$$

and

$$\widehat{P}(s, a, \mu) := \arg\min_{P \in P(s,a,\mu)} \sum_{s' \in S} P(s')\big(r(s, a, s', \mu) + \gamma v_\mu(s')\big). \tag{75}$$

For $(\pi, p) \in \Pi \times \mathcal{K}$, define the induced Markov kernel on $\mathcal{S}$ as

$$K_{\pi,p}(s'|s) := \sum_{a \in A} \pi(a|s)\, p(s'|s, a). \tag{76}$$

Now define three correspondences:

$$\mathsf{BR}(\mu) := \Big\{\pi \in \Pi : \operatorname{supp} \pi(\cdot|s) \subset D(s, \mu)\ \forall s \in S\Big\}, \tag{77}$$

$$\mathsf{WC}(\mu) := \Big\{p \in \mathcal{K} : p(\cdot|s, a) \in \widehat{P}(s, a, \mu)\ \forall (s, a) \in S \times A\Big\}, \tag{78}$$

$$\mathsf{Inv}(\pi, p) := \Big\{\eta \in \Delta(S) : \eta K_{\pi,p} = \eta\Big\}. \tag{79}$$

Finally set

$$\Phi(\mu, \pi, p) := \mathsf{Inv}(\pi, p) \times \mathsf{BR}(\mu) \times \mathsf{WC}(\mu). \tag{80}$$

A.2.1 Preparatory lemmas

Lemma 6 (Continuity of the robust Bellman data)**.** *Under Assumption 1:*

(i) for every fixed $v \in \mathbb{R}^S$, the map $(\mu, s, a) \mapsto (Q_\mu v)(s, a)$ is continuous;

(ii) the map $\mu \mapsto v_\mu$ is continuous in $\|\cdot\|_\infty$;

(iii) the map $\mu \mapsto Q_\mu(s, a)$ is continuous for every $(s, a) \in S \times A$.

Proof. For fixed v, define

$$H_{s,a}(\mu, P; v) := \sum_{s' \in S} P(s')\big(r(s,a,s',\mu) + \gamma v(s')\big). \tag{81}$$

By boundedness and continuity of r in μ, $H_{s,a}$ is continuous in (μ, P). Since $P(s,a,\mu)$ is compact-valued and continuous as a correspondence, Berge's maximum theorem implies that

$$(\mu, s, a) \longmapsto (Q_\mu v)(s,a) = \min_{P \in P(s,a,\mu)} H_{s,a}(\mu, P; v) \tag{82}$$

is continuous.

Now let $\mu_n \to \mu$. Since $v_{\mu_n} = T_{\mu_n} v_{\mu_n}$ and $v_\mu = T_\mu v_\mu$,

$$\|v_{\mu_n} - v_\mu\|_\infty \le \|T_{\mu_n} v_{\mu_n} - T_{\mu_n} v_\mu\|_\infty + \|T_{\mu_n} v_\mu - T_\mu v_\mu\|_\infty \le \gamma \|v_{\mu_n} - v_\mu\|_\infty + \|T_{\mu_n} v_\mu - T_\mu v_\mu\|_\infty. \tag{83}$$

Hence

$$(1-\gamma)\|v_{\mu_n} - v_\mu\|_\infty \le \|T_{\mu_n} v_\mu - T_\mu v_\mu\|_\infty \to 0, \tag{84}$$

because the first part of the lemma gives continuity of $T_\mu v_\mu$ in μ. Therefore $v_{\mu_n} \to v_\mu$.

Finally,

$$Q_\mu(s,a) = \min_{P \in P(s,a,\mu)} \sum_{s'} P(s')\big(r(s,a,s',\mu) + \gamma v_\mu(s')\big), \tag{85}$$

so the continuity of $\mu \mapsto v_\mu$ and Berge's theorem again imply the continuity of $\mu \mapsto Q_\mu(s,a)$. □

Lemma 7 (Properties of the fixed-point correspondence)**.** *Under Assumption 1:*

(i) $\mathsf{BR}(\mu)$*,* $\mathsf{WC}(\mu)$*, and* $\mathsf{Inv}(\pi, p)$ *are all nonempty, compact, and convex;*

(ii) the graphs of BR*,* WC*, and* Inv *are closed;*

(iii) therefore Φ *has nonempty compact convex values and closed graph.*

Proof. Step 1: $\mathsf{BR}(\mu)$. For each state $\mathcal{S}$, $D(s,\mu) \neq \varnothing$ because $\mathcal{A}$ is finite. The set of distributions on $\mathcal{A}$ supported on $D(s,\mu)$ is a simplex, hence nonempty, compact, and convex. Taking the finite product over $\mathcal{S}$ gives the same properties for $\mathsf{BR}(\mu)$.

To prove closed graph, suppose $\mu_n \to \mu$, $\pi_n \to \pi$, and $\pi_n \in \mathsf{BR}(\mu_n)$. Fix $s \in S$ and $a \in A$ with $\pi(a|s) > 0$. For n large enough one has $\pi_n(a|s) > 0$, hence $a \in D(s, \mu_n)$. By Lemma 6, $Q_{\mu_n}(s,\cdot) \to Q_\mu(s,\cdot)$ uniformly over the finite set $\mathcal{A}$. Therefore

$$Q_\mu(s,a) = \lim_{n\to\infty} Q_{\mu_n}(s,a) = \lim_{n\to\infty} \max_{b \in A} Q_{\mu_n}(s,b) = \max_{b \in A} Q_\mu(s,b), \tag{86}$$

so $a \in D(s,\mu)$. Thus $\pi \in \mathsf{BR}(\mu)$.

Step 2: $\mathsf{WC}(\mu)$. For each (s,a), the objective defining $\widehat{P}(s,a,\mu)$ is continuous and affine in P, while $P(s,a,\mu)$ is nonempty, compact, and convex. Hence $\widehat{P}(s,a,\mu)$ is nonempty, compact, and convex. Taking the finite product over (s,a) gives the same properties for $\mathsf{WC}(\mu)$.

To prove closed graph, suppose $\mu_n \to \mu$, $p_n \to p$, and $p_n \in \mathsf{WC}(\mu_n)$. Fix (s,a). By the closed-graph property of the ambiguity correspondence, $p(\cdot|s,a) \in P(s,a,\mu)$. Let $q \in P(s,a,\mu)$ be arbitrary. By lower hemicontinuity of $P(s,a,\cdot)$, there are $q_n \in P(s,a,\mu_n)$ such that $q_n \to q$. Since $p_n(\cdot|s,a)$ minimizes the robust Bellman objective at μ_n,

$$\sum_{s'} p_n(s'|s,a)\big(r(s,a,s',\mu_n) + \gamma v_{\mu_n}(s')\big) \le \sum_{s'} q_n(s')\big(r(s,a,s',\mu_n) + \gamma v_{\mu_n}(s')\big). \tag{87}$$

Passing to the limit and using Lemma 6 gives

$$\sum_{s'} p(s'|s,a)\big(r(s,a,s',\mu) + \gamma v_\mu(s')\big) \le \sum_{s'} q(s')\big(r(s,a,s',\mu) + \gamma v_\mu(s')\big). \tag{88}$$

Since q was arbitrary in $P(s,a,\mu)$, $p(\cdot|s,a) \in \widehat{P}(s,a,\mu)$. Hence $p \in \mathsf{WC}(\mu)$.

Step 3: $\mathsf{Inv}(\pi, p)$. For fixed (π, p), the map

$$\eta \longmapsto \eta K_{\pi,p} \tag{89}$$

is continuous from $\Delta(S)$ into itself. Because $\Delta(S)$ is compact and convex, Brouwer's fixed point theorem gives at least one invariant distribution. The set $\mathsf{Inv}(\pi, p)$ is the solution set of the linear equation $\eta K_{\pi,p} = \eta$ inside $\Delta(S)$, hence it is compact and convex.

For closed graph, let $(\pi_n, p_n, \eta_n) \to (\pi, p, \eta)$ with $\eta_n \in \mathsf{Inv}(\pi_n, p_n)$. Since

$$\eta_n(s') = \sum_{s\in S} \eta_n(s) \sum_{a\in A} \pi_n(a|s) p_n(s'|s, a) \qquad \forall s' \in S, \tag{90}$$

passing to the limit in these finite sums yields $\eta K_{\pi,p} = \eta$. Hence $\eta \in \mathsf{Inv}(\pi, p)$.

Step 4: product correspondence. The product of correspondences with nonempty compact convex values and closed graph has the same properties, so Φ does as well. □

A.2.2 Existence

Theorem 6. *Under Assumption 1, there exists a stationary robust mean-field equilibrium.*

Proof. The space

$$X := \Delta(S) \times \Pi \times \mathcal{K} \tag{91}$$

is nonempty, compact, and convex in a finite-dimensional Euclidean space. By Lemma 7, the correspondence $\Phi : X \rightrightarrows X$ has nonempty compact convex values and closed graph. Hence Kakutani's fixed point theorem gives a triple $(\mu^\star, \pi^\star, p^\star) \in X$ such that

$$(\mu^\star, \pi^\star, p^\star) \in \Phi(\mu^\star, \pi^\star, p^\star). \tag{92}$$

From the definition of Φ, the fixed-point relation means:

(i) $\mu^\star \in \mathsf{Inv}(\pi^\star, p^\star)$, i.e.

$$\mu^\star(s') = \sum_{s\in S} \mu^\star(s) \sum_{a\in A} \pi^\star(a|s) p^\star(s'|s, a) \qquad \forall s' \in S; \tag{93}$$

(ii) $\pi^\star \in \mathsf{BR}(\mu^\star)$, so $\operatorname{supp} \pi^\star(\cdot|s) \subset D(s, \mu^\star)$ for every state $\mathcal{S}$;

(iii) $p^\star \in \mathsf{WC}(\mu^\star)$, so $p^\star(\cdot|s, a) \in \widehat{P}(s, a, \mu^\star)$ for every (s, a).

Now apply Proposition 1 at the population $\mu^\star$. Because $\pi^\star$ is pointwise optimal at every state and $p^\star$ is a pointwise minimizing kernel at every state-action pair, Proposition 1 yields

$$V_{\mu^\star} = \inf_p J_{\mu^\star}(\pi^\star, p) = J_{\mu^\star}(\pi^\star, p^\star). \tag{94}$$

Together with item (i), this is exactly a stationary robust mean-field equilibrium. □

B Approximation of the N-player robust games

B.1 N-player robust games and DRMGs

We first note that, our N-player robust game is a robust Markov game on the enlarged state space $\mathcal{S}^N$, with

- joint state s^N,
- joint action a^N,
- Player i's reward $r_i(s^N, a^N, (s')^N) = r\big(s^i, a^i, (s')^i, e^N(s^N)\big)$,

with a (s, a)-rectangular uncertainty set $\mathfrak{P}^N(s^N, a^N)$. Thus, our N-player robust game is a special robust Markov game on the joint state space $\mathcal{S}^N$ with mean-field coupling.

Notably, the objective function ((10)) in N-player robust games is defined as

$$J_i^N(\pi^N) = \inf_{(p_t^N)} \mathbb{E}^{\pi^N,(p_t^N)}\Big[\sum_{t\geq 0}\gamma^t\, r(s_t^i, a_t^i, s_{t+1}^i, e^N(s_t^N))\Big], \tag{95}$$

where each agent concerns its own worst-case performance under the uncertainty set. Although in (10), the worst-case kernel is taken w.r.t. non-stationary kernel sequence (p_t), we will show the worst-case is achieved at a stationary kernel. Thus, (10) matches the robust value function of the standard robust Markov game.

B.2 Counterexample

Theorem 7. *Fix $\gamma \in (0,1)$ and $\kappa > 1/\gamma$, and set*

$$\varepsilon_0 \;:=\; \frac{\gamma\kappa - 1}{1-\gamma} \;>\; 0.$$

There exist a robust mean-field game satisfying Assumption 1 and a stationary robust mean-field equilibrium $(\mu^\star, \pi^\star, p^\star)$ of it such that:

- *(a) Assumption 2 fails for the constant deviation sequence $\pi^{(N)} \equiv \pi^\star$ under* every *choice of minimizing laws $P^{N|(N)}$; and*
- *(b) for* every *$N \in \mathbb{N}$, the symmetric profile $\pi^{N|\star}$ is not an ε-Nash equilibrium of the N-player robust game for any $\varepsilon < \varepsilon_0$.*

Proof. **The model.** Let $\mathcal{S} = \{0,1\}$, $\mathcal{A} = \{b, o\}$, $\mathfrak{P}(s,a,\mu) = \Delta(\mathcal{S})$ for all (s,a,μ) (full ambiguity), and

$$r(s,a,s',\mu) \;=\; \begin{cases} 1 - \kappa\,\mu(1), & a = b, \\ 0, & a = o. \end{cases}$$

Assumption 1 holds: rewards are bounded and affine (hence continuous) in μ; the constant correspondence $\mu \mapsto \Delta(\mathcal{S})$ is nonempty, convex, compact valued, with closed graph, and lower hemicontinuous.

Step 1: $(\mu^\star, \pi^\star, p^\star) = (\delta_0,\ \pi^\star(b|\cdot) \equiv 1,\ p^\star(\cdot|s,a) \equiv \delta_0)$ **is a stationary robust MFE.** At the population $\mu^\star = \delta_0$ we have $\mu^\star(1) = 0$, so $r(s,b,s',\mu^\star) = 1$ and $r(s,o,s',\mu^\star) = 0$ for all s, s'. Since the reward does not depend on the successor state, for every stationary policy π and every admissible stationary kernel p,

$$J_{\mu^\star}(\pi, p) \;=\; \sum_{t\geq 0}\gamma^t\, P^{\pi,p}(a_t = b) \;\leq\; \frac{1}{1-\gamma},$$

with equality when $\pi = \pi^\star$, for *every* p. Hence

$$V_{\mu^\star} \;=\; \sup_\pi \inf_p J_{\mu^\star}(\pi, p) \;=\; \frac{1}{1-\gamma} \;=\; \inf_p J_{\mu^\star}(\pi^\star, p) \;=\; J_{\mu^\star}(\pi^\star, p^\star),$$

which verifies Definition 2(i)–(ii). Consistency (Definition 2(iii)) holds because $K_{\pi^\star,p^\star}(\cdot|s) = \delta_0$ for every s, so $\delta_0 K_{\pi^\star,p^\star} = \delta_0 = \mu^\star$.

Step 2: Assumption 2 fails under every choice of minimizing laws. Take the constant deviation sequence $\pi^{(N)} \equiv \pi^\star$, so $\pi^{N|(N)} = \pi^{N|\star}$ and every player always plays b. Fix N and let $(p_t^N)_{t\geq 0}$ be *any* admissible kernel sequence for the N-player robust game. Since $s_0^i \sim \mu^\star = \delta_0$ i.i.d., we have $e_N(s_0^N) = \delta_0$ almost surely, and the realized reward of player 1 at time t is $1 - \kappa\, e_N(s_t^N)(1)$. Therefore the payoff under (p_t^N) equals

$$\sum_{t\geq 0}\gamma^t\Big(1 - \kappa\,\mathbb{E}\big[e_N(s_t^N)(1)\big]\Big) \;=\; \frac{1}{1-\gamma} \;-\; \kappa\sum_{t\geq 1}\gamma^t\,\mathbb{E}\big[e_N(s_t^N)(1)\big]. \tag{96}$$

Because $e_N(\cdot)(1) \in [0,1]$ pointwise, (96) is bounded below by

$$\frac{1}{1-\gamma} - \kappa\sum_{t\geq 1}\gamma^t \;=\; \frac{1-\gamma\kappa}{1-\gamma},$$

with equality if and only if $\mathbb{E}[e_N(s_t^N)(1)] = 1$ for every $t \geq 1$, i.e., if and only if $s_t^i = 1$ almost surely for all players i and all $t \geq 1$. The lower bound is attained by the admissible constant product kernel $p^N(\cdot \mid s^N, a^N) := \delta_1 \otimes \cdots \otimes \delta_1 \in \mathfrak{P}^N(s^N, a^N)$, which sends every player to state 1 at time 1 and keeps them there. Hence

$$J_1^N(\pi^{N|\star}) \;=\; \frac{1-\gamma\kappa}{1-\gamma}, \tag{97}$$

and, crucially, *every* law $P^{N|(N)}$ attaining the infimum satisfies $s_1^1 = 1$ almost surely. Consequently, under *every* minimizing law,

$$Q_0^{N|(N)} \;=\; \delta_{(0,\,b,\,1,\,\delta_0)}, \qquad \text{whereas} \qquad Q_0^{\star(N)} \;=\; \delta_{(0,\,b,\,0,\,\delta_0)}$$

(the proxy chain has $s_0 = 0$, $a_0 = b$, $s_1 \sim p^\star(\cdot|0, b) = \delta_0$, population coordinate $\mu^\star = \delta_0$). These are Dirac masses at points $z \neq z'$ of Z with $d_Z(z, z') = \mathbf{1}\{(0, b, 1) \neq (0, b, 0)\} + \|\delta_0 - \delta_0\|_1 = 1$, so

$$W_1\big(Q_0^{N|(N)},\, Q_0^{\star(N)}\big) \;=\; 1 \qquad \text{for every } N \text{ and every choice of minimizing laws.}$$

Thus (16) fails at $t = 0$ for this deviation sequence regardless of how the minimizing laws are chosen, i.e., Assumption 2 does not hold. This proves (a).

Step 3: uniform Nash gap. Consider the unilateral deviation $\pi(o|s) \equiv 1$ by player 1. Its reward is identically 0 along every trajectory, regardless of the kernel sequence, so $J_1^N(\pi^{N|\star,-1} \oplus \pi) = 0$ for every N. Combining with (97) and $\kappa > 1/\gamma$,

$$\sup_{\pi'} J_1^N(\pi^{N|\star,-1} \oplus \pi') \;-\; J_1^N(\pi^{N|\star}) \;\geq\; 0 - \frac{1-\gamma\kappa}{1-\gamma} \;=\; \frac{\gamma\kappa - 1}{1-\gamma} \;=\; \varepsilon_0 \;>\; 0$$

for every N. Hence $\pi^{N|\star}$ is not an ε-Nash equilibrium for any $\varepsilon < \varepsilon_0$, proving (b). □

B.3 Auxiliary facts

Lemma 8 (Finite-player robust dynamic programming)**.** *Fix $N \in \mathbb{N}$, a profile $\pi^N \in \Pi^N$, and an agent $i \in \{1, \ldots, N\}$. Define the one-step reward on $S^N \times A^N \times S^N$ by*

$$\rho_i(s^N, a^N, s'^N) := r\big(s^i, a^i, (s')^i, e^N(s^N)\big).$$

Then the following hold.

1. *There exists a unique function $u^{N,i} : S^N \to \mathbb{R}$ satisfying*

$$u^{N,i}(s^N) = \sum_{a^N \in A^N} \pi^N(a^N \mid s^N) \min_{p \in P^N(s^N, a^N)} \sum_{s'^N \in S^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big).$$

2. *The robust payoff in* (95) *satisfies*

$$J_i^N(\pi^N) = \mathbb{E}_{s_0^N \sim (\mu^\star)^{\otimes N}}\big[u^{N,i}(s_0^N)\big].$$

3. *The infimum in* (95) *is attained by a stationary kernel*

$$p^{N,i,\star} : S^N \times A^N \to \Delta(S^N)$$

with

$$p^{N,i,\star}(\cdot \mid s^N, a^N) \in P^N(s^N, a^N) \qquad \forall (s^N, a^N) \in S^N \times A^N.$$

More precisely, one may choose $p^{N,i,\star}$ so that

$$\sum_{s'^N} p^{N,i,\star}(s'^N \mid s^N, a^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big) = \min_{p \in P^N(s^N, a^N)} \sum_{s'^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big),$$

and, if we use the constant sequence $p_t^N \equiv p^{N,i,\star}$, then

$$J_i^N(\pi^N) = \mathbb{E}^{\pi^N, p^{N,i,\star}}\left[\sum_{t=0}^{\infty} \gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N)\right].$$

Proof. For $u \in \mathbb{R}^{S^N}$, define the Bellman operator $T^{N,i} : \mathbb{R}^{S^N} \to \mathbb{R}^{S^N}$ by

$$(T^{N,i}u)(s^N) := \sum_{a^N \in A^N} \pi^N(a^N \mid s^N) \min_{p \in P^N(s^N,a^N)} \sum_{s'^N \in S^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u(s'^N)\Big).$$

By Lemma 9, each set $P^N(s^N, a^N)$ is nonempty and compact, so every inner minimum is attained.

We first prove (i). Fix $u, v \in \mathbb{R}^{S^N}$ and $s^N \in S^N$. For each $a^N \in A^N$, define

$$\Psi_{a^N}(u) := \min_{p \in P^N(s^N,a^N)} \sum_{s'^N \in S^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u(s'^N)\Big).$$

Choose

$$p_u \in \arg\min_{p \in P^N(s^N,a^N)} \sum_{s'^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u(s'^N)\Big).$$

Then

$$\begin{aligned}\Psi_{a^N}(u) - \Psi_{a^N}(v) &\le \sum_{s'^N} p_u(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u(s'^N)\Big) \\ &\quad - \sum_{s'^N} p_u(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma v(s'^N)\Big) \\ &= \gamma \sum_{s'^N} p_u(s'^N)\big(u(s'^N) - v(s'^N)\big) \le \gamma\|u - v\|_\infty.\end{aligned}$$

Exchanging the roles of u and v gives

$$|\Psi_{a^N}(u) - \Psi_{a^N}(v)| \le \gamma\|u - v\|_\infty.$$

Averaging with respect to $\pi^N(\cdot \mid s^N)$ yields

$$|(T^{N,i}u)(s^N) - (T^{N,i}v)(s^N)| \le \gamma\|u - v\|_\infty.$$

Taking the supremum over $s^N \in S^N$, we obtain

$$\|T^{N,i}u - T^{N,i}v\|_\infty \le \gamma\|u - v\|_\infty.$$

Since $\gamma \in (0, 1)$, $T^{N,i}$ is a contraction on the complete metric space $(\mathbb{R}^{S^N}, \|\cdot\|_\infty)$. By Banach's fixed-point theorem, there exists a unique fixed point $u^{N,i}$ such that

$$u^{N,i} = T^{N,i}u^{N,i}.$$

This proves (i).

We next prove (ii) and (iii). For each $(s^N, a^N) \in S^N \times A^N$, choose

$$p^{N,i,\star}(\cdot \mid s^N, a^N) \in \arg\min_{p \in P^N(s^N,a^N)} \sum_{s'^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big).$$

Because $S^N \times A^N$ is finite, this defines a stationary selector with no measurability issue. By construction,

$$\sum_{s'^N} p^{N,i,\star}(s'^N \mid s^N, a^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big) = \min_{p \in P^N(s^N,a^N)} \sum_{s'^N} p(s'^N)\Big(\rho_i(s^N, a^N, s'^N) + \gamma u^{N,i}(s'^N)\Big). \tag{98}$$

Fix an initial state $x^N \in S^N$, and consider the controlled Markov chain generated by the stationary profile π^N and the stationary kernel $p^{N,i,\star}$, started from $s_0^N = x^N$. Using the fixed-point identity $u^{N,i} = T^{N,i}u^{N,i}$ together with (98), we get

$$u^{N,i}(x^N) = \mathbb{E}_{x^N}^{\pi^N, p^{N,i,\star}}\Big[\rho_i(s_0^N, a_0^N, s_1^N) + \gamma u^{N,i}(s_1^N)\Big].$$

Iterating this identity yields, for every $T \ge 1$,

$$u^{N,i}(x^N) = \mathbb{E}_{x^N}^{\pi^N, p^{N,i,\star}}\left[\sum_{t=0}^{T-1} \gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N) + \gamma^T u^{N,i}(s_T^N)\right].$$

Since $|\rho_i| \le \|r\|_\infty$, the fixed point is bounded:

$$\|u^{N,i}\|_\infty = \|T^{N,i}u^{N,i}\|_\infty \le \|r\|_\infty + \gamma\|u^{N,i}\|_\infty,$$

hence

$$\|u^{N,i}\|_\infty \le \frac{\|r\|_\infty}{1-\gamma}.$$

Therefore $\gamma^T u^{N,i}(s_T^N) \to 0$ in L^1, and letting $T \to \infty$ gives

$$u^{N,i}(x^N) = \mathbb{E}_{x^N}^{\pi^N, p^{N,i,\star}}\left[\sum_{t=0}^{\infty}\gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N)\right].$$

Averaging over $x^N = s_0^N \sim (\mu^\star)^{\otimes N}$, we obtain

$$\mathbb{E}^{\pi^N, p^{N,i,\star}}\left[\sum_{t=0}^{\infty}\gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N)\right] = \mathbb{E}_{s_0^N\sim(\mu^\star)^{\otimes N}}\left[u^{N,i}(s_0^N)\right].$$

So the stationary selector $p^{N,i,\star}$ achieves the value $\mathbb{E}[u^{N,i}(s_0^N)]$.

It remains to show that no admissible nonstationary sequence can do better for the minimizing player. Let $(q_t^N)_{t\ge0}$ be any admissible sequence with

$$q_t^N(\cdot \mid s^N, a^N) \in P^N(s^N, a^N) \qquad \forall t,\ (s^N, a^N).$$

Consider the controlled process under π^N and $(q_t^N)_{t\ge0}$. For every state s^N and every time t, by the definition of the minimum in $T^{N,i}$,

$$\begin{aligned} u^{N,i}(s^N) &= \sum_{a^N}\pi^N(a^N \mid s^N) \min_{p\in P^N(s^N,a^N)}\sum_{y^N} p(y^N)\Big(\rho_i(s^N, a^N, y^N) + \gamma u^{N,i}(y^N)\Big) \\ &\le \sum_{a^N}\pi^N(a^N \mid s^N)\sum_{y^N} q_t^N(y^N \mid s^N, a^N)\Big(\rho_i(s^N, a^N, y^N) + \gamma u^{N,i}(y^N)\Big). \end{aligned}$$

Thus, along the controlled process,

$$u^{N,i}(s_t^N) \le \mathbb{E}^{\pi^N,(q_t^N)}\Big[\rho_i(s_t^N, a_t^N, s_{t+1}^N) + \gamma u^{N,i}(s_{t+1}^N) \,\Big|\, s_t^N\Big].$$

Taking expectations, multiplying by γ^t, and summing from $t = 0$ to $T-1$, we get the telescoping estimate

$$\mathbb{E}^{\pi^N,(q_t^N)}\left[\sum_{t=0}^{T-1}\gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N)\right] \ge \mathbb{E}^{\pi^N,(q_t^N)}\big[u^{N,i}(s_0^N)\big] - \mathbb{E}^{\pi^N,(q_t^N)}\big[\gamma^T u^{N,i}(s_T^N)\big].$$

Again, boundedness of $u^{N,i}$ implies the last term tends to 0 as $T \to \infty$, so

$$\mathbb{E}^{\pi^N,(q_t^N)}\left[\sum_{t=0}^{\infty}\gamma^t \rho_i(s_t^N, a_t^N, s_{t+1}^N)\right] \ge \mathbb{E}_{s_0^N\sim(\mu^\star)^{\otimes N}}\left[u^{N,i}(s_0^N)\right].$$

Since $(q_t^N)_{t\ge0}$ was arbitrary, every admissible sequence yields a payoff at least $\mathbb{E}[u^{N,i}(s_0^N)]$, while the stationary selector $p^{N,i,\star}$ attains exactly this value. Therefore

$$J_i^N(\pi^N) = \mathbb{E}_{s_0^N\sim(\mu^\star)^{\otimes N}}\left[u^{N,i}(s_0^N)\right],$$

and the infimum is attained by the stationary kernel $p^{N,i,\star}$. This proves (ii) and (iii). □

Fix a stationary robust mean-field equilibrium $(\mu^\star, \pi^\star, p^\star)$. Define the equilibrium proxy value

$$V^\star := \sum_{s\in S}\mu^\star(s)\, v_{\mu^\star}(s). \tag{99}$$

By Lemma 5 applied at $\mu = \mu^\star$,

$$V^\star = \sup_{\pi: S\to\Delta(A)} \mathbb{E}_{s_0\sim\mu^\star}^{\pi, p^\star}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t, s_{t+1}, \mu^\star)\right], \tag{100}$$

and the supremum is attained at $\pi = \pi^\star$.

Lemma 9. *Assume Assumption 1. Then for every N and (s^N, a^N), the set $\mathfrak{P}^N(s^N, a^N)$ defined in (7) is nonempty and compact.*

Proof. Non-emptiness: choose for each i some $p^i(\cdot \mid s^i, a^i) \in \mathfrak{P}(s^i, a^i, e^N(s^N))$ and take the product.

Compactness: each $\mathfrak{P}(s^i, a^i, e^N(s^N))$ is compact in the simplex $\Delta(\mathcal{S})$. Their finite product is compact in $(\Delta(\mathcal{S}))^N$. The map $(p^1, \dots, p^N) \mapsto \bigotimes_{i=1}^N p^i$ is continuous (finite products of coordinates), hence the image $\mathfrak{P}^N(s^N, a^N)$ is compact. □

Lemma 10. *Let X be finite and Y Polish. Let $\lambda_n \in \mathfrak{P}(X)$ and kernels $K_n(\cdot \mid x) \in \mathfrak{P}(Y)$. If $\lambda_n \to \lambda$ and $K_n(\cdot \mid x) \to K(\cdot \mid x)$ for each $x \in X$, then the measures $\Lambda_n(dx, dy) := \lambda_n(dx)K_n(dy \mid x)$ satisfy $\Lambda_n \to \Lambda(dx, dy) := \lambda(dx)K(dy \mid x)$.*

Proof. Since X is finite, weak convergence of λ_n is pointwise convergence of masses. For bounded continuous $g : X \times Y \to \mathbb{R}$,

$$\int g \, d\Lambda_n = \sum_{x \in X} \lambda_n(\{x\}) \int_Y g(x, y) \, K_n(dy \mid x). \tag{101}$$

For each x, $y \mapsto g(x, y)$ is bounded continuous, so the inner integrals converge by $K_n(\cdot \mid x) \to K(\cdot \mid x)$, and the coefficients converge by $\lambda_n(\{x\}) \to \lambda(\{x\})$. Sum is finite, so the limits pass through the sum. □

B.4 Convergence of discounted payoffs

Lemma 11. *Let (Z, d_Z) be a compact metric space and $f : Z \to \mathbb{R}$ continuous, with modulus of continuity $\omega_f(\delta) := \sup\{|f(z) - f(z')| : d_Z(z, z') \le \delta\}$. Then for all $\mu, \nu \in \Delta(Z)$ and all $\delta > 0$,*

$$\Big| \int_Z f \, d\mu - \int_Z f \, d\nu \Big| \;\le\; \omega_f(\delta) \;+\; \frac{2 \, \|f\|_\infty}{\delta} \, W_1(\mu, \nu). \tag{102}$$

Consequently, if $(\mu_N)_N, (\nu_N)_N \subset \Delta(Z)$ satisfy $W_1(\mu_N, \nu_N) \to 0$, then $\int f \, d\mu_N - \int f \, d\nu_N \to 0$.

Proof. Since Z is compact, f is bounded and uniformly continuous, so $\omega_f(\delta) < \infty$ for all $\delta > 0$ and $\omega_f(\delta) \downarrow 0$ as $\delta \downarrow 0$. Because Z is a compact Polish space, the infimum defining $W_1(\mu, \nu)$ is attained by some coupling $\Gamma \in \Delta(Z \times Z)$ with marginals μ, ν. Writing $(Z, Z') \sim \Gamma$,

$$\Big| \int f \, d\mu - \int f \, d\nu \Big| \le \mathbb{E}_\Gamma \big| f(Z) - f(Z') \big| \le \omega_f(\delta) + 2 \, \|f\|_\infty \; P_\Gamma\big(d_Z(Z, Z') > \delta\big) \le \omega_f(\delta) + \frac{2 \, \|f\|_\infty}{\delta} \, \mathbb{E}_\Gamma \, d_Z(Z, Z'),$$

where the last step is Markov's inequality; since $\mathbb{E}_\Gamma d_Z(Z, Z') = W_1(\mu, \nu)$, (102) follows. For the convergence claim, take $\limsup_N$ in (102) and then let $\delta \downarrow 0$. □

Proposition 4. *Under Assumptions 1 and 2, for every deviation sequence $(\pi^{(N)})_N$,*

$$\lim_{N \to \infty} \Big(J_1^N\big(\pi^{N|\star,-1} \oplus \pi^{(N)}\big) \;-\; J^\star\big(\pi^{(N)}\big) \Big) \;=\; 0,$$

where $J^\star(\pi^{(N)})$ is the proxy payoff.

Proof. Define $\bar{r} : Z \to \mathbb{R}$ by $\bar{r}(s, a, s', \nu) := r(s, a, s', \nu)$. By Assumption 1(1) it is bounded by $\|r\|_\infty$; it is continuous on Z because $\mathcal{S}, \mathcal{A}$ are finite (so the first three coordinates are discrete) and $\nu \mapsto r(s, a, s', \nu)$ is continuous for each fixed (s, a, s') by Assumption 1(3). The space $Z = (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \times \Delta(\mathcal{S})$ is compact.

Let $P^{N|(N)}$ be minimizing laws as in Assumption 2. Since $|r| \le \|r\|_\infty$ and $\gamma \in (0, 1)$, Fubini's theorem gives the absolutely convergent expansions

$$J_1^N\big(\pi^{N|\star,-1} \oplus \pi^{(N)}\big) = \sum_{t \ge 0} \gamma^t \int_Z \bar{r} \, dQ_t^{N|(N)}, \qquad J^\star\big(\pi^{(N)}\big) = \sum_{t \ge 0} \gamma^t \int_Z \bar{r} \, dQ_t^{\star(N)}.$$

Set $\Delta_{N,t} := \int \bar{r} \, dQ_t^{N|(N)} - \int \bar{r} \, dQ_t^{\star(N)}$. For each fixed t, Assumption 2 gives $W_1(Q_t^{N|(N)}, Q_t^{\star(N)}) \to 0$ as $N \to \infty$, so Lemma 11 applied with $f = \bar{r}$ yields $\Delta_{N,t} \to 0$. Moreover $|\Delta_{N,t}| \le 2 \, \|r\|_\infty$ uniformly in (N, t), and $\sum_t \gamma^t \cdot 2 \, \|r\|_\infty < \infty$; dominated convergence (over t, with respect to the summable envelope $2 \, \|r\|_\infty \gamma^t$) gives $\sum_t \gamma^t \Delta_{N,t} \to 0$, which is the claim. □

B.5 Approximate Nash equilibrium

Theorem 8. *Assume Assumptions 1 and 2. Let*

$$\pi^{N|\star} := (\pi^\star, \dots, \pi^\star) \in \Pi^N. \tag{103}$$

Then, for every $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that, for all $N \geq N(\varepsilon)$, the profile $\pi^{N|\star}$ is an ε-Nash equilibrium.

Proof. By symmetry of the N-player robust game, it suffices to prove the ε-best-response inequality for player 1.

Step 1: equilibrium payoff converges to $V^\star$. Apply Proposition 4 to the constant deviation sequence $\pi^{(N)} \equiv \pi^\star$. Then $\pi^{N|(N)} = \pi^{N|\star}$ for every N, and the proxy chain coincides with the Markov chain driven by $(\mu^\star, \pi^\star, p^\star)$. Hence

$$\lim_{N\to\infty} J_1^N(\pi^{N|\star}) = V^\star. \tag{104}$$

Step 2: every deviation sequence has proxy payoff at most $V^\star$. Fix any deviation sequence $(\pi^{(N)})_{N\in\mathbb{N}} \subset \Pi$ and set $\pi^{N|(N)} = (\pi^{(N)}, \pi^\star, \dots, \pi^\star)$. By Proposition 4,

$$J_1^N(\pi^{N|(N)}) - \mathbb{E}_{P^{*(N)}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu^\star)\right] \to 0. \tag{105}$$

For each N, the proxy payoff on the right-hand side is bounded above by $V^\star$, because $V^\star$ is the optimal value of the fixed-kernel MDP with kernel $p^\star$ and initial law $\mu^\star$. Therefore

$$\limsup_{N\to\infty} J_1^N(\pi^{N|(N)}) \leq V^\star. \tag{106}$$

Step 3: contradiction argument. Assume, toward a contradiction, that the conclusion is false. Then there exists $\varepsilon > 0$ and a subsequence $(N_k)_{k\in\mathbb{N}}$ such that, for each k, one can find a unilateral deviation policy $\widehat{\pi}^{(N_k)} \in \Pi$ satisfying

$$J_1^{N_k}(\pi^{N_k|\star}) + \varepsilon < J_1^{N_k}\big(\pi^{N_k|\star,-1} \oplus \widehat{\pi}^{(N_k)}\big). \tag{107}$$

Extend these deviations to a full sequence $(\pi^{(N)})_{N\in\mathbb{N}}$ by setting

$$\pi^{(N)} = \begin{cases} \widehat{\pi}^{(N_k)}, & N = N_k \text{ for some } k, \\ \pi^\star, & \text{otherwise.} \end{cases} \tag{108}$$

Then

$$J_1^{N_k}(\pi^{N_k|\star}) + \varepsilon < J_1^{N_k}(\pi^{N_k|(N_k)}) \qquad \forall\, k. \tag{109}$$

Taking $\limsup_{k\to\infty}$ and using Steps 1 and 2 yields

$$V^\star + \varepsilon \leq \limsup_{k\to\infty} J_1^{N_k}(\pi^{N_k|(N_k)}) \leq V^\star, \tag{110}$$

a contradiction.

Therefore, for every $\varepsilon > 0$, there exists $N(\varepsilon)$ such that

$$J_1^N(\pi^{N|\star}) + \varepsilon \geq \sup_{\pi\in\Pi} J_1^N(\pi^{N|\star,-1} \oplus \pi) \qquad \forall\, N \geq N(\varepsilon). \tag{111}$$

By symmetry, the same bound holds for every player, so $\pi^{N|\star}$ is an ε-Nash equilibrium for all sufficiently large N. □

C Non-Asymptotic Finite-N Approximation via Proxy Comparison

This appendix proves the finite-N result stated in Section 6.3. The key point is that the comparison is made between the actual N-player robust game and the proxy chain driven by the equilibrium kernel $p^\star$.

C.1 Metrics and one-step laws

Let

$$X := \mathcal{S} \times \mathcal{A} \times \mathcal{S}, \qquad Z := X \times \Delta(\mathcal{S}). \tag{112}$$

We equip X with the discrete metric

$$d_X(x, \tilde{x}) := \mathbf{1}_{\{x \neq \tilde{x}\}}, \tag{113}$$

and Z with

$$d_Z\big((x, \nu), (\tilde{x}, \tilde{\nu})\big) := d_X(x, \tilde{x}) + \|\nu - \tilde{\nu}\|_1 = \mathbf{1}_{\{x \neq \tilde{x}\}} + \|\nu - \tilde{\nu}\|_1. \tag{114}$$

We write W_1 for the corresponding Wasserstein-1 distance on $\mathcal{P}(Z)$.

Fix $N \in \mathbb{N}$ and a unilateral deviation policy $\pi \in \Pi$. Let $P^{N,\pi}$ be a minimizing law attaining

$$J_1^N(\pi^{N|\star,-1} \oplus \pi), \tag{115}$$

whose existence follows from Lemma 8, and define

$$Q_t^{N,\pi} := \mathrm{Law}_{P^{N,\pi}}\big(s_t^1, a_t^1, s_{t+1}^1, e^N(s_t^N)\big), \qquad t \in \mathbb{N}_0. \tag{116}$$

Define the proxy chain P^π by

$$s_0 \sim \mu^\star, \qquad a_t \sim \pi(\cdot \mid s_t), \qquad s_{t+1} \sim p^\star(\cdot \mid s_t, a_t), \tag{117}$$

and let

$$Q_t^\pi := \mathrm{Law}_{P^\pi}\big(s_t, a_t, s_{t+1}, \mu^\star\big), \qquad t \in \mathbb{N}_0. \tag{118}$$

Throughout this appendix, we assume Assumption 3 and the following Assumption 6 (which generalizes the Lipschitz Assumption 4).

Assumption 6 (Hölder reward regularity)**.** *There exist $L_r > 0$ and $\alpha \in (0, 1]$ such that, for all $s \in S$, $a \in A$, $s' \in S$, and all $\nu, \tilde{\nu} \in \Delta(S)$,*

$$|r(s, a, s', \nu) - r(s, a, s', \tilde{\nu})| \leq L_r \|\nu - \tilde{\nu}\|_1^\alpha.$$

C.2 Discussion of Assumption 3

Assumption 3 should be viewed as a quantitative robust propagation-of-chaos / Nash-certainty-equivalence hypothesis, extended from non-robust MFGs. In a standard non-robust mean-field model, analogous estimates are often obtained from regularity of the dynamics together with concentration of the empirical measure. In the present robust setting, however, the minimizing kernel in the N-player game may exploit the degrees of freedom of the remaining $N - 1$ players to alter the empirical distribution in a way that does not vanish automatically as $N \to \infty$. For this reason, the baseline compactness and continuity assumptions used for existence do not imply Assumption 3; indeed, the counterexample from the asymptotic subsection already shows that even qualitative convergence may fail without an additional stabilization property.

At the same time, Assumption 3 is substantially weaker than imposing a concrete structural sufficient condition inside the theorem. It is stated directly at the level of the observable one-step law and is agnostic to the mechanism producing that law: no product-form realization of the minimizing kernel, no particular selector for the worst-case transition, and no contractivity property of the robust population map is built into the statement. Any model-specific coupling, concentration, or perturbation argument that yields the bound in Assumption 3 can therefore be substituted directly into the theorem without changing the payoff-comparison proof.

The assumption is also weaker than a trajectory-level approximation requirement. We do not ask for a single coupling of the entire paths of the N-player game and the proxy chain. Instead, we only require control of the one-step law at each time t, because this is exactly the object that appears in the discounted sum of stage rewards. The array $(\delta_{N,t})$ is allowed to depend on t, which permits moderate accumulation of finite-N errors before discounting. The theorem only needs

$$\sum_{t \geq 0} \gamma^t \Big(2\|r\|_\infty \delta_{N,t} + L_r \delta_{N,t}^\alpha\Big) < \infty.$$

Finally, the uniformity over unilateral deviations is essential for the ε_N-Nash conclusion. If the approximation bound were available only for a fixed deviation policy, then one would obtain only a deviation-dependent comparison estimate, not a single Nash-gap bound after taking the supremum over all deviations. In many stable regimes one expects $\delta_{N,t}$ to exhibit the usual $N^{-1/2}$ scaling, possibly multiplied by a factor with mild growth in t, but we do not encode any such verification mechanism into the theorem because the purpose of the theorem is to isolate the weakest quantitative input needed for the finite-N comparison.

C.3 Proxy-payoff comparison under Hölder rewards

Proposition 5. *Assume Assumptions 1, 3, and 6. Fix $N \in \mathbb{N}$ and a unilateral deviation policy $\pi \in \Pi$. Then*

$$\left| J_1^N(\pi^{N|\star,-1} \oplus \pi) - \mathbb{E}_{P^\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu^\star)\right]\right| \le \sum_{t=0}^{\infty} \gamma^t \Big(2\|r\|_\infty \, \delta_{N,t} + L_r \, \delta_{N,t}^\alpha\Big). \tag{119}$$

Proof. Fix N and π. Since r is bounded, Tonelli's theorem yields

$$J_1^N(\pi^{N|\star,-1} \oplus \pi) = \sum_{t=0}^{\infty} \gamma^t \int r \, dQ_t^{N,\pi}, \tag{120}$$

and

$$\mathbb{E}_{P^\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu^\star)\right] = \sum_{t=0}^{\infty} \gamma^t \int r \, dQ_t^{\pi}. \tag{121}$$

Fix $t \in \mathbb{N}_0$, and let Γ_t^π be an optimal coupling of $Q_t^{N,\pi}$ and Q_t^π. Write

$$(Z_t^N, Z_t^\pi) \sim \Gamma_t^\pi, \tag{122}$$

with

$$Z_t^N = (X_t^N, \nu_t^N) \in X \times \Delta(\mathcal{S}), \qquad Z_t^\pi = (X_t^\pi, \mu^\star) \in X \times \Delta(\mathcal{S}). \tag{123}$$

Then

$$\begin{aligned}\left| \int r \, dQ_t^{N,\pi} - \int r \, dQ_t^{\pi} \right| &\le \int \left| r(X_t^N, \nu_t^N) - r(X_t^\pi, \mu^\star) \right| \, d\Gamma_t^\pi \\ &\le 2\|r\|_\infty \, \Gamma_t^\pi(X_t^N \neq X_t^\pi) + L_r \int \|\nu_t^N - \mu^\star\|_1^\alpha \, d\Gamma_t^\pi.\end{aligned}$$

Since

$$\mathbf{1}_{\{X_t^N \neq X_t^\pi\}} \le d_Z(Z_t^N, Z_t^\pi), \qquad \|\nu_t^N - \mu^\star\|_1 \le d_Z(Z_t^N, Z_t^\pi), \tag{124}$$

we obtain

$$\Gamma_t^\pi(X_t^N \neq X_t^\pi) \le \int d_Z \, d\Gamma_t^\pi = W_1(Q_t^{N,\pi}, Q_t^\pi) \le \delta_{N,t}. \tag{125}$$

Moreover, by concavity of $x \mapsto x^\alpha$ on $[0, \infty)$,

$$\int \|\nu_t^N - \mu^\star\|_1^\alpha \, d\Gamma_t^\pi \le \left(\int \|\nu_t^N - \mu^\star\|_1 \, d\Gamma_t^\pi\right)^\alpha \le \left(\int d_Z \, d\Gamma_t^\pi\right)^\alpha \le \delta_{N,t}^\alpha. \tag{126}$$

Hence

$$\left| \int r \, dQ_t^{N,\pi} - \int r \, dQ_t^{\pi} \right| \le 2\|r\|_\infty \, \delta_{N,t} + L_r \, \delta_{N,t}^\alpha. \tag{127}$$

Multiplying by γ^t and summing over $t \ge 0$ proves the claim. □

C.4 Finite-N ε_N-Nash bound

Theorem 9. *Assume Assumptions 1, 3, and 6. Let $(\mu^\star, \pi^\star, p^\star)$ be a stationary robust mean-field equilibrium and let*

$$\pi^{N|\star} := (\pi^\star, \dots, \pi^\star) \in \Pi^N. \tag{128}$$

Then $\pi^{N|\star}$ is an ε_N-Nash equilibrium, where

$$\varepsilon_N := 2\sum_{t=0}^{\infty} \gamma^t \Big(2\|r\|_\infty \, \delta_{N,t} + L_r \, \delta_{N,t}^\alpha\Big). \tag{129}$$

Proof. Define

$$B_N := \sum_{t=0}^{\infty} \gamma^t \Big(2\|r\|_\infty \, \delta_{N,t} + L_r \, \delta_{N,t}^{\alpha} \Big). \tag{130}$$

Also define the equilibrium proxy value

$$V^\star := \sup_{\pi : S \to \Delta(A)} \mathbb{E}^{\pi, p^\star}_{s_0 \sim \mu^\star} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu^\star) \right]. \tag{131}$$

By Lemma 5, the supremum is attained at $\pi = \pi^\star$.

Apply Proposition 5 with $\pi = \pi^\star$. Then

$$J_1^N(\pi^{N|\star}) \geq V^\star - B_N. \tag{132}$$

Now fix any unilateral deviation policy $\pi \in \Pi$. Applying Proposition 5 to this π gives

$$J_1^N(\pi^{N|\star,-1} \oplus \pi) \leq \mathbb{E}_{P^\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}, \mu^\star) \right] + B_N \leq V^\star + B_N. \tag{133}$$

Therefore,

$$J_1^N(\pi^{N|\star,-1} \oplus \pi) - J_1^N(\pi^{N|\star}) \leq 2B_N = 2 \sum_{t=0}^{\infty} \gamma^t \Big(2\|r\|_\infty \, \delta_{N,t} + L_r \, \delta_{N,t}^{\alpha} \Big). \tag{134}$$

Taking the supremum over $\pi \in \Pi$ gives the desired best-response bound for player 1. By symmetry of the game, the same estimate holds for every player, and thus $\pi^{N|\star}$ is an ε_N-Nash equilibrium. □

Corollary 1 (Rate under $\delta_{N,t} \leq C_\delta(1+t)/\sqrt{N}$)**.** *Assume the hypotheses of Theorem 9, and suppose that there exists $C_\delta > 0$ such that*

$$\delta_{N,t} \leq \frac{C_\delta(1+t)}{\sqrt{N}} \qquad \forall\, N \in \mathbb{N},\ t \in \mathbb{N}_0. \tag{135}$$

Then

$$\varepsilon_N \leq \frac{4\|r\|_\infty C_\delta}{\sqrt{N}} \sum_{t=0}^{\infty} \gamma^t (1+t) + \frac{2L_r C_\delta^\alpha}{N^{\alpha/2}} \sum_{t=0}^{\infty} \gamma^t (1+t)^\alpha, \tag{136}$$

and hence

$$\varepsilon_N = O\Big(N^{-\alpha/2} \Big). \tag{137}$$

Proof. Under the assumed estimate on $\delta_{N,t}$,

$$2\|r\|_\infty \, \delta_{N,t} \leq \frac{2\|r\|_\infty C_\delta(1+t)}{\sqrt{N}}, \qquad L_r \, \delta_{N,t}^{\alpha} \leq \frac{L_r C_\delta^\alpha (1+t)^\alpha}{N^{\alpha/2}}. \tag{138}$$

Insert these bounds into the formula for ε_N from Theorem 9. Since

$$\sum_{t=0}^{\infty} \gamma^t (1+t) < \infty, \qquad \sum_{t=0}^{\infty} \gamma^t (1+t)^\alpha < \infty, \tag{139}$$

the stated estimate follows. Because $\alpha \in (0, 1]$, the slower-decaying term is $N^{-\alpha/2}$, which gives the displayed rate. □

Corollary 2 (Lipschitz specialization)**.** *Assume Assumption 3, and assume, instead of Assumption 6, that*

$$|r(s, a, s', \nu) - r(s, a, s', \tilde{\nu})| \leq L_r \|\nu - \tilde{\nu}\|_1 \qquad \forall\, s, a, s', \nu, \tilde{\nu}. \tag{140}$$

Then $\pi^{N|\star}$ is an $\varepsilon_N^{\mathrm{Lip}}$-Nash equilibrium with

$$\varepsilon_N^{\mathrm{Lip}} = 2(2\|r\|_\infty + L_r) \sum_{t=0}^{\infty} \gamma^t \delta_{N,t}. \tag{141}$$

If, in addition,

$$\delta_{N,t} \le \frac{C_\delta(1+t)}{\sqrt{N}} \qquad \forall\, N \in \mathbb{N},\ t \in \mathbb{N}_0, \tag{142}$$

then

$$\varepsilon_N^{\mathrm{Lip}} \le \frac{2(2\|r\|_\infty + L_r)C_\delta}{(1-\gamma)^2\sqrt{N}}. \tag{143}$$

Proof. Set $\alpha = 1$ in Theorem 9. Then

$$\varepsilon_N = 2\sum_{t=0}^{\infty} \gamma^t \big(2\|r\|_\infty + L_r\big)\delta_{N,t} = 2(2\|r\|_\infty + L_r)\sum_{t=0}^{\infty}\gamma^t \delta_{N,t}, \tag{144}$$

which proves the first statement. If, moreover, $\delta_{N,t} \le C_\delta(1+t)/\sqrt{N}$, then

$$\varepsilon_N^{\mathrm{Lip}} \le \frac{2(2\|r\|_\infty + L_r)C_\delta}{\sqrt{N}} \sum_{t=0}^{\infty}\gamma^t(1+t). \tag{145}$$

Since

$$\sum_{t=0}^{\infty}\gamma^t(1+t) = \frac{1}{(1-\gamma)^2}, \tag{146}$$

the explicit bound follows. □

C.5 A concrete sufficient regime

The non-asymptotic theorem is stated under the law-level proxy-comparison assumption. In this section, we aim to verify that assumption in a concrete stable regime.

For a unilateral deviation policy $\pi \in \Pi$ and $N \in \mathbb{N}$, let

$$\pi^{N,\pi} := (\pi, \pi^\star, \dots, \pi^\star) \in \Pi^N.$$

Assumption 7 (A sufficient stable regime)**.** *There exist a selector $\bar{p} : S \times A \times \Delta(S) \to \Delta(S)$ and constants $L_P \ge 0$ and $\rho_{\mathrm{mix}} \in [0,1)$ such that:*

1. *For every $(s,a,\nu) \in S \times A \times \Delta(S)$,*
$$\bar{p}(\cdot \mid s,a,\nu) \in P(s,a,\nu).$$

2. *The selector is Lipschitz in the population argument:*
$$\max_{(s,a)\in S\times A} \big\|\bar{p}(\cdot \mid s,a,\nu) - \bar{p}(\cdot \mid s,a,\tilde{\nu})\big\|_1 \le L_P\|\nu - \tilde{\nu}\|_1, \qquad \forall \nu,\tilde{\nu} \in \Delta(S).$$

3. *For each $\nu \in \Delta(S)$, define the $\pi^\star$-controlled kernel*
$$K_\nu(s' \mid s) := \sum_{a\in A} \pi^\star(a \mid s)\,\bar{p}(s' \mid s,a,\nu).$$
Its Dobrushin coefficient satisfies
$$\alpha(K_\nu) := \max_{s,\tilde{s}\in S} d_{\mathrm{TV}}\big(K_\nu(\cdot \mid s), K_\nu(\cdot \mid \tilde{s})\big) \le \rho_{\mathrm{mix}}, \qquad \forall \nu \in \Delta(S).$$

4. *The selector is compatible with the mean-field worst-case kernel at equilibrium:*
$$p^\star(\cdot \mid s,a) = \bar{p}(\cdot \mid s,a,\mu^\star), \qquad \forall (s,a) \in S \times A.$$

5. *For every $N \in \mathbb{N}$ and every unilateral deviation policy $\pi \in \Pi$, there exists an admissible minimizing kernel sequence $(p_t^{N,\pi,\star})_{t\ge 0}$ attaining the infimum in the definition of $J_1^N(\pi^{N,\pi})$ such that for every $t \ge 0$ and every $(s^N,a^N) \in S^N \times A^N$,*
$$p_t^{N,\pi,\star}(\cdot \mid s^N,a^N) = \bigotimes_{i=1}^{N} \bar{p}(\cdot \mid s^i,a^i,e_N(s^N)).$$

Finally, assume

$$\rho := \rho_{\mathrm{mix}} + L_P < 1.$$

Assumption 7 is strong, but it is only used to justify a benchmark rate for the quantitative proxy comparison. The product-form realization in part (5) is precisely what prevents the minimizing N-player kernel from introducing additional cross-agent dependence beyond the empirical distribution.

For fixed N and $\pi \in \Pi$, let $\mathbb{P}^{N,\pi}$ denote the law induced by the profile $\pi^{N,\pi}$ and the minimizing kernel sequence from Assumption 7(5). Let

$$\mu_t^N := e_N(s_t^N).$$

Let $\mathbb{P}^{\pi}$ denote the proxy chain defined by

$$s_0 \sim \mu^\star, \qquad a_t \sim \pi(\cdot \mid s_t), \qquad s_{t+1} \sim p^\star(\cdot \mid s_t, a_t),$$

and write

$$Q_t^{N,\pi} := \mathrm{Law}_{\mathbb{P}^{N,\pi}}(s_t^1, a_t^1, s_{t+1}^1, \mu_t^N), \qquad Q_t^{\pi} := \mathrm{Law}_{\mathbb{P}^{\pi}}(s_t, a_t, s_{t+1}, \mu^\star).$$

Lemma 12 (Contraction of the robust population map)**.** *Under Assumption 7, the map*

$$F(\nu) := \nu K_\nu, \qquad \nu \in \Delta(S),$$

satisfies

$$\|F(\nu) - F(\tilde\nu)\|_1 \le \rho \, \|\nu - \tilde\nu\|_1, \qquad \forall \nu, \tilde\nu \in \Delta(S).$$

In particular, F is a contraction on $(\Delta(S), \|\cdot\|_1)$.

Proof. Fix $\nu, \tilde\nu \in \Delta(S)$. Add and subtract $\nu K_{\tilde\nu}$:

$$\|F(\nu) - F(\tilde\nu)\|_1 = \|\nu K_\nu - \tilde\nu K_{\tilde\nu}\|_1 \le \|\nu K_\nu - \nu K_{\tilde\nu}\|_1 + \|\nu K_{\tilde\nu} - \tilde\nu K_{\tilde\nu}\|_1.$$

For the first term,

$$\|\nu K_\nu - \nu K_{\tilde\nu}\|_1 \le \max_{s \in S} \|K_\nu(\cdot \mid s) - K_{\tilde\nu}(\cdot \mid s)\|_1.$$

Moreover, for each $s \in S$,

$$\begin{aligned}
\|K_\nu(\cdot \mid s) - K_{\tilde\nu}(\cdot \mid s)\|_1 &= \left\| \sum_{a \in A} \pi^\star(a \mid s) \Big(\bar p(\cdot \mid s, a, \nu) - \bar p(\cdot \mid s, a, \tilde\nu) \Big) \right\|_1 \\
&\le \sum_{a \in A} \pi^\star(a \mid s) \, \big\| \bar p(\cdot \mid s, a, \nu) - \bar p(\cdot \mid s, a, \tilde\nu) \big\|_1 \\
&\le L_P \|\nu - \tilde\nu\|_1.
\end{aligned}$$

Hence

$$\|\nu K_\nu - \nu K_{\tilde\nu}\|_1 \le L_P \|\nu - \tilde\nu\|_1.$$

For the second term, the standard Dobrushin contraction inequality gives

$$\|\nu K_{\tilde\nu} - \tilde\nu K_{\tilde\nu}\|_1 \le \alpha(K_{\tilde\nu}) \, \|\nu - \tilde\nu\|_1 \le \rho_{\mathrm{mix}} \|\nu - \tilde\nu\|_1.$$

Combining the two bounds yields

$$\|F(\nu) - F(\tilde\nu)\|_1 \le (L_P + \rho_{\mathrm{mix}}) \|\nu - \tilde\nu\|_1 = \rho \|\nu - \tilde\nu\|_1.$$

□

Lemma 13 (Uniform empirical-distribution error)**.** *Under Assumption 7, for every $N \in \mathbb{N}$, every unilateral deviation policy $\pi \in \Pi$, and every $t \ge 0$,*

$$\mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big] \le \frac{2\sqrt{|S|}}{(1-\rho)\sqrt{N}} + \frac{2}{(1-\rho)N}.$$

Consequently,

$$\sup_{t \ge 0} \mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big] \le \frac{C_\mu}{\sqrt{N}}, \qquad C_\mu := \frac{2\sqrt{|S|} + 2}{1-\rho}.$$

The bound is uniform over the unilateral deviation policy π.

Proof. Fix N and π, and write

$$a_t := \mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big].$$

Let

$$\mathcal{F}_t := \sigma(s_0^N, \dots, s_t^N)$$

be the state filtration up to time t. By Assumption 7(5), conditional on $\mathcal{F}_t$ the next states $s_{t+1}^1, \dots, s_{t+1}^N$ are independent, and their conditional laws are

$$\mathbb{P}^{N,\pi}(s_{t+1}^1 \in \cdot \mid \mathcal{F}_t) = \sum_{a \in A} \pi(a \mid s_t^1)\, \bar{p}(\cdot \mid s_t^1, a, \mu_t^N),$$

and, for $i \geq 2$,

$$\mathbb{P}^{N,\pi}(s_{t+1}^i \in \cdot \mid \mathcal{F}_t) = \sum_{a \in A} \pi^\star(a \mid s_t^i)\, \bar{p}(\cdot \mid s_t^i, a, \mu_t^N) = K_{\mu_t^N}(\cdot \mid s_t^i).$$

Define the two random probability vectors

$$G_t(\cdot) := \sum_{a \in A} \pi(a \mid s_t^1)\, \bar{p}(\cdot \mid s_t^1, a, \mu_t^N), \qquad H_t(\cdot) := K_{\mu_t^N}(\cdot \mid s_t^1).$$

Then

$$\begin{aligned}
\mathbb{E}^{N,\pi}[\mu_{t+1}^N \mid \mathcal{F}_t] &= \frac{1}{N} G_t + \frac{1}{N} \sum_{i=2}^{N} K_{\mu_t^N}(\cdot \mid s_t^i) \\
&= \mu_t^N K_{\mu_t^N} + \frac{1}{N}\big(G_t - H_t\big) \\
&= F(\mu_t^N) + \frac{1}{N}\big(G_t - H_t\big).
\end{aligned}$$

Set

$$\xi_{t+1} := \mu_{t+1}^N - \mathbb{E}^{N,\pi}[\mu_{t+1}^N \mid \mathcal{F}_t].$$

We first bound the fluctuation term. For any $s' \in S$,

$$\mu_{t+1}^N(s') = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{s_{t+1}^i = s'\}}.$$

Conditional on $\mathcal{F}_t$, the indicators on the right-hand side are independent Bernoulli random variables, hence

$$\mathrm{Var}\big(\mu_{t+1}^N(s') \mid \mathcal{F}_t\big) \leq \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{P}^{N,\pi}(s_{t+1}^i = s' \mid \mathcal{F}_t).$$

Summing over $s' \in S$ gives

$$\sum_{s' \in S} \mathrm{Var}\big(\mu_{t+1}^N(s') \mid \mathcal{F}_t\big) \leq \frac{1}{N}.$$

Therefore, by Jensen and Cauchy–Schwarz,

$$\begin{aligned}
\mathbb{E}^{N,\pi}[\|\xi_{t+1}\|_1 \mid \mathcal{F}_t] &\leq \sqrt{|S|}\, \mathbb{E}^{N,\pi}[\|\xi_{t+1}\|_2 \mid \mathcal{F}_t] \\
&\leq \sqrt{|S|}\, \Big(\mathbb{E}^{N,\pi}[\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t]\Big)^{1/2} \\
&= \sqrt{|S|}\, \left(\sum_{s' \in S} \mathrm{Var}\big(\mu_{t+1}^N(s') \mid \mathcal{F}_t\big)\right)^{1/2} \\
&\leq \sqrt{\frac{|S|}{N}}.
\end{aligned}$$

Next, by the triangle inequality,

$$\|\mu_{t+1}^N - \mu^\star\|_1 \leq \|\xi_{t+1}\|_1 + \|F(\mu_t^N) - F(\mu^\star)\|_1 + \frac{1}{N}\|G_t - H_t\|_1.$$

By Assumption 7(4) and the consistency condition of the mean-field equilibrium,

$$F(\mu^\star) = \mu^\star.$$

Also, $G_t, H_t \in \Delta(S)$, so $\|G_t - H_t\|_1 \le 2$. Taking expectations and using Lemma 12 yields

$$a_{t+1} \le \rho a_t + \sqrt{\frac{|S|}{N}} + \frac{2}{N}.$$

Iterating this recursion gives

$$a_t \le \rho^t a_0 + \frac{1-\rho^t}{1-\rho}\left(\sqrt{\frac{|S|}{N}} + \frac{2}{N}\right).$$

It remains to bound a_0. Since μ_0^N is the empirical distribution of N i.i.d. samples from $\mu^\star$, the same variance argument as above gives

$$a_0 = \mathbb{E}^{N,\pi}[\|\mu_0^N - \mu^\star\|_1] \le \sqrt{\frac{|S|}{N}}.$$

Hence

$$\begin{aligned} a_t &\le \sqrt{\frac{|S|}{N}} + \frac{1}{1-\rho}\left(\sqrt{\frac{|S|}{N}} + \frac{2}{N}\right) \\ &= \frac{2-\rho}{1-\rho}\sqrt{\frac{|S|}{N}} + \frac{2}{(1-\rho)N} \\ &\le \frac{2\sqrt{|S|}}{(1-\rho)\sqrt{N}} + \frac{2}{(1-\rho)N}. \end{aligned}$$

The stated uniform $C_\mu/\sqrt{N}$ bound follows because $N^{-1} \le N^{-1/2}$ for $N \ge 1$. □

Proposition 6 (One-step law comparison)**.** *Under Assumption 7, for every $N \in \mathbb{N}$, every unilateral deviation policy $\pi \in \Pi$, and every $t \ge 0$,*

$$W_1\big(Q_t^{N,\pi}, Q_t^\pi\big) \le \mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big] + L_P \sum_{k=0}^{t} \mathbb{E}^{N,\pi}\big[\|\mu_k^N - \mu^\star\|_1\big].$$

In particular,

$$W_1\big(Q_t^{N,\pi}, Q_t^\pi\big) \le \frac{C_\mu}{\sqrt{N}}\big(1 + L_P(t+1)\big).$$

Proof. Fix N, π, and t. We construct a coupling between

$$Z_t^{N,\pi} := (s_t^1, a_t^1, s_{t+1}^1, \mu_t^N) \qquad \text{and} \qquad Z_t^\pi := (s_t, a_t, s_{t+1}, \mu^\star).$$

Start with $s_0 = s_0^1$, where $s_0^1 \sim \mu^\star$.

At time t, given (s_t^1, s_t), proceed as follows.

1. If $s_t^1 = s_t$, sample a single action $a_t \sim \pi(\cdot \mid s_t)$ and set $a_t^1 := a_t$.
2. If $s_t^1 \ne s_t$, couple $a_t^1 \sim \pi(\cdot \mid s_t^1)$ and $a_t \sim \pi(\cdot \mid s_t)$ arbitrarily.

Thus,

$$\{s_t^1 = s_t\} \subseteq \{a_t^1 = a_t\}.$$

Next, conditional on $(s_t^1, a_t^1, s_t, a_t, \mu_t^N)$,

- under $\mathbb{P}^{N,\pi}$, the first player's next-state law is $\bar{p}(\cdot \mid s_t^1, a_t^1, \mu_t^N)$, by Assumption 7(5);
- under $\mathbb{P}^\pi$, the proxy next-state law is $p^\star(\cdot \mid s_t, a_t) = \bar{p}(\cdot \mid s_t, a_t, \mu^\star)$, by Assumption 7(4).

Whenever $(s_t^1, a_t^1) = (s_t, a_t)$, couple s_{t+1}^1 and s_{t+1} by a maximal coupling of

$$\bar{p}(\cdot \mid s_t, a_t, \mu_t^N) \qquad \text{and} \qquad \bar{p}(\cdot \mid s_t, a_t, \mu^\star).$$

When $(s_t^1, a_t^1) \neq (s_t, a_t)$, couple s_{t+1}^1 and s_{t+1} arbitrarily.

Define

$$e_t := \mathbb{P}(s_t^1 \neq s_t)$$

under this coupling. Then $e_0 = 0$, and by the union bound,

$$\begin{aligned} e_{t+1} &\leq \mathbb{P}(s_t^1 \neq s_t) + \mathbb{P}\big(s_t^1 = s_t,\ a_t^1 = a_t,\ s_{t+1}^1 \neq s_{t+1}\big) \\ &\leq e_t + \mathbb{E}\Big[d_{\mathrm{TV}}\big(\bar{p}(\cdot \mid s_t, a_t, \mu_t^N), \bar{p}(\cdot \mid s_t, a_t, \mu^\star)\big)\Big] \\ &\leq e_t + \frac{L_P}{2}\,\mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big]. \end{aligned}$$

By induction,

$$e_t \leq \frac{L_P}{2}\sum_{k=0}^{t-1}\mathbb{E}^{N,\pi}\big[\|\mu_k^N - \mu^\star\|_1\big], \qquad e_t + e_{t+1} \leq L_P \sum_{k=0}^{t}\mathbb{E}^{N,\pi}\big[\|\mu_k^N - \mu^\star\|_1\big].$$

Now recall that the metric on $Z = (S \times A \times S) \times \Delta(S)$ is

$$d_Z\big((x,\nu),(\tilde{x},\tilde{\nu})\big) = \mathbf{1}_{\{x \neq \tilde{x}\}} + \|\nu - \tilde{\nu}\|_1.$$

Therefore,

$$d_Z\big(Z_t^{N,\pi}, Z_t^\pi\big) = \mathbf{1}_{\{(s_t^1, a_t^1, s_{t+1}^1) \neq (s_t, a_t, s_{t+1})\}} + \|\mu_t^N - \mu^\star\|_1.$$

By construction, if both $s_t^1 = s_t$ and $s_{t+1}^1 = s_{t+1}$ occur, then necessarily $a_t^1 = a_t$, and hence

$$\mathbf{1}_{\{(s_t^1, a_t^1, s_{t+1}^1) \neq (s_t, a_t, s_{t+1})\}} \leq \mathbf{1}_{\{s_t^1 \neq s_t\}} + \mathbf{1}_{\{s_{t+1}^1 \neq s_{t+1}\}}.$$

Taking expectations gives

$$\begin{aligned} \mathbb{E}\big[d_Z(Z_t^{N,\pi}, Z_t^\pi)\big] &\leq e_t + e_{t+1} + \mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big] \\ &\leq L_P \sum_{k=0}^{t}\mathbb{E}^{N,\pi}\big[\|\mu_k^N - \mu^\star\|_1\big] + \mathbb{E}^{N,\pi}\big[\|\mu_t^N - \mu^\star\|_1\big]. \end{aligned}$$

Since $W_1(Q_t^{N,\pi}, Q_t^\pi)$ is the infimum of this expected cost over all couplings, the first inequality follows. The second inequality is immediate from Lemma 13. □

Corollary 3. *Under Assumption 7, it holds with*

$$\delta_{N,t} := \frac{C_\mu}{\sqrt{N}}\big(1 + L_P(t+1)\big), \qquad C_\mu := \frac{2\sqrt{|S|} + 2}{1 - \rho}.$$

Proof. This is exactly Proposition 6. □

Corollary 4 (Resulting finite-N Nash-gap bound)**.** *Assume, in addition, the α-Hölder reward condition; namely, for some $\alpha \in (0,1]$ and $L_r \geq 0$,*

$$|r(s,a,s',\nu) - r(s,a,s',\tilde{\nu})| \leq L_r\|\nu - \tilde{\nu}\|_1^\alpha, \qquad \forall s,a,s',\nu,\tilde{\nu}.$$

Let $R_\infty := \|r\|_\infty$. Then the Nash-gap bound satisfies

$$\varepsilon_N \leq \frac{4R_\infty C_\mu}{\sqrt{N}}\left(\frac{1}{1-\gamma} + \frac{L_P}{(1-\gamma)^2}\right) + \frac{2L_r C_\mu^\alpha}{N^{\alpha/2}}\left(\frac{1}{1-\gamma} + \frac{L_P^\alpha}{(1-\gamma)^2}\right).$$

In particular,

$$\varepsilon_N = O\big(N^{-\alpha/2}\big).$$

When $\alpha = 1$ (the Lipschitz case), this becomes

$$\varepsilon_N \leq \frac{2(2R_\infty + L_r)C_\mu}{\sqrt{N}}\left(\frac{1}{1-\gamma} + \frac{L_P}{(1-\gamma)^2}\right).$$

Proof. By Corollary 3, Theorem 9 applies with

$$\delta_{N,t} = \frac{C_\mu}{\sqrt{N}}\big(1 + L_P(t+1)\big).$$

Using Theorem 9,

$$\varepsilon_N \le 2\sum_{t\ge 0}\gamma^t\Big(2R_\infty\delta_{N,t} + L_r\delta_{N,t}^\alpha\Big).$$

For the linear term,

$$\sum_{t\ge 0}\gamma^t\delta_{N,t} = \frac{C_\mu}{\sqrt{N}}\sum_{t\ge 0}\gamma^t\big(1 + L_P(t+1)\big) = \frac{C_\mu}{\sqrt{N}}\left(\frac{1}{1-\gamma} + \frac{L_P}{(1-\gamma)^2}\right).$$

For the Hölder term, since $\alpha \in (0,1]$ and $(x+y)^\alpha \le x^\alpha + y^\alpha$ for $x, y \ge 0$,

$$\big(1 + L_P(t+1)\big)^\alpha \le 1 + L_P^\alpha(t+1)^\alpha \le 1 + L_P^\alpha(t+1).$$

Hence

$$\begin{aligned}\sum_{t\ge 0}\gamma^t\delta_{N,t}^\alpha &\le \frac{C_\mu^\alpha}{N^{\alpha/2}}\sum_{t\ge 0}\gamma^t\Big(1 + L_P^\alpha(t+1)\Big)\\ &= \frac{C_\mu^\alpha}{N^{\alpha/2}}\left(\frac{1}{1-\gamma} + \frac{L_P^\alpha}{(1-\gamma)^2}\right).\end{aligned}$$

Substituting the last two estimates into the theorem yields the stated bound. The Lipschitz specialization is the case $\alpha = 1$. □

D Algorithmic solutions to Robust MFGs

D.1 Discussion: verifying the Lipschitz-mixing assumptions

(1) Verifying mixing via a Doeblin/minorization condition. A convenient sufficient condition for (34) is: there exist $\eta \in (0,1]$ and $\psi \in \Delta(\mathcal{S})$ such that for all μ, all $s \in S$,

$$K_\mu(\cdot \mid s) \ge \eta\,\psi(\cdot) \quad \text{(componentwise)}. \tag{147}$$

Then $\alpha(K_\mu) \le 1 - \eta$ uniformly over μ.

Thus, as long as the kernel K_μ has a full support set of $\mathcal{S}$, (i.e., every entry of K_μ is positive), the condition will be satisfied. We highlight that this condition can be satisfied by distributional uncertainty set with small radius, and is also considered in standard robust RL literature, e.g., [Wang et al., 2023a, Chen et al., 2026].

(2) Bounding L_K by policy and nature sensitivities. From $K_\mu(\cdot \mid s) = \sum_a \pi_\mu(a \mid s)p_\mu(\cdot \mid s, a)$, one can bound, for each s,

$$\|K_\mu(\cdot \mid s) - K_{\tilde\mu}(\cdot \mid s)\|_1 \le \underbrace{\max_a \|p_\mu(\cdot \mid s,a) - p_{\tilde\mu}(\cdot \mid s,a)\|_1}_{\text{worst-case kernel sensitivity}} + \underbrace{\|\pi_\mu(\cdot \mid s) - \pi_{\tilde\mu}(\cdot \mid s)\|_1}_{\text{policy sensitivity}}, \tag{148}$$

since $\|p_{\tilde\mu}(\cdot \mid s, a)\|_1 = 1$. Hence a sufficient condition for (35) is the existence of Lipschitz selections $\mu \mapsto p_\mu$ and $\mu \mapsto \pi_\mu$.

We then verify (or relax) the kernel Lipschitz condition (35) in Assumption 5 for several standard ambiguity correspondences. We focus on the ambiguity families: total-variation balls and Wasserstein balls.

Notation. Fix $(s,a) \in S \times A$. For a population distribution $\mu \in \Delta(\mathcal{S})$ and a bounded value function $v \in \mathbb{R}^S$, define the one-step *cost vector*

$$c_{\mu,v}^{s,a}(s') \;:=\; r(s,a,s',\mu) + \gamma v(s'), \qquad s' \in S. \tag{149}$$

For a fixed selection rule $\mu \mapsto (\pi_\mu, p_\mu)$, recall $K_\mu(\cdot \mid s) = \sum_a \pi_\mu(a \mid s)p_\mu(\cdot \mid s, a)$.

D.1.1 From policy/kernel Lipschitzness to L_K

Lemma 14 (A convenient bound for L_K)**.** *Assume there exist constants $L_\pi, L_p \geq 0$ such that for all $\mu, \tilde{\mu} \in \Delta(\mathcal{S})$,*

$$\max_{s \in S} \|\pi_\mu(\cdot \mid s) - \pi_{\tilde{\mu}}(\cdot \mid s)\|_1 \leq L_\pi \|\mu - \tilde{\mu}\|_1, \tag{150}$$

$$\max_{(s,a) \in S \times A} \|p_\mu(\cdot \mid s, a) - p_{\tilde{\mu}}(\cdot \mid s, a)\|_1 \leq L_p \|\mu - \tilde{\mu}\|_1. \tag{151}$$

Then the induced kernel K_μ satisfies (35) *with*

$$L_K \leq L_\pi + L_p. \tag{152}$$

Proof. Fix $\mu, \tilde{\mu}$ and $s \in S$. Add and subtract $\sum_a \pi_\mu(a \mid s)\, p_{\tilde{\mu}}(\cdot \mid s, a)$:

$$\|K_\mu(\cdot \mid s) - K_{\tilde{\mu}}(\cdot \mid s)\|_1 \leq \Big\| \sum_a \pi_\mu(a \mid s) \big(p_\mu(\cdot \mid s, a) - p_{\tilde{\mu}}(\cdot \mid s, a)\big) \Big\|_1 + \Big\| \sum_a \big(\pi_\mu(a \mid s) - \pi_{\tilde{\mu}}(a \mid s)\big) p_{\tilde{\mu}}(\cdot \mid s, a) \Big\|_1. \tag{153}$$

The first term is at most $\max_a \|p_\mu(\cdot \mid s, a) - p_{\tilde{\mu}}(\cdot \mid s, a)\|_1$. The second term is at most $\|\pi_\mu(\cdot \mid s) - \pi_{\tilde{\mu}}(\cdot \mid s)\|_1$ since $\|p_{\tilde{\mu}}(\cdot \mid s, a)\|_1 = 1$. Taking maxima over s and using (150)–(151) yields (152). □

Policy Lipschitzness: gap vs. softmax. In general, argmax-based policies can be discontinuous in μ due to ties. Two standard sufficient conditions to obtain (150) are: (i) *local action-gap/uniqueness* near $\mu^\star$ (then $L_\pi = 0$ locally), or (ii) *entropy-regularized policies* (softmax), which are globally Lipschitz.

Lemma 15 (Softmax policy is Lipschitz in Q)**.** *Fix $\tau > 0$. For each $s \in S$ define*

$$\pi^\tau(a \mid s; Q) \;:=\; \frac{\exp(Q(s,a)/\tau)}{\sum_{b \in A} \exp(Q(s,b)/\tau)}. \tag{154}$$

Then for any two $Q, \tilde{Q}$,

$$\|\pi^\tau(\cdot \mid s; Q) - \pi^\tau(\cdot \mid s; \tilde{Q})\|_1 \leq \frac{2}{\tau} \max_{a \in A} |Q(s,a) - \tilde{Q}(s,a)|. \tag{155}$$

Consequently, if $\max_{s,a} |Q_\mu(s,a) - Q_{\tilde{\mu}}(s,a)| \leq L_Q \|\mu - \tilde{\mu}\|_1$, then (150) *holds with $L_\pi \leq 2L_Q/\tau$.*

Proof. Let $u(a) = Q(s,a)/\tau$ and $\tilde{u}(a) = \tilde{Q}(s,a)/\tau$. The softmax map $u \mapsto \mathrm{softmax}(u)$ has Jacobian $J(u) = \mathrm{diag}(\pi) - \pi\pi^\top$ whose operator norm from ℓ_∞ to ℓ_1 is at most 2. Thus $\|\pi(u) - \pi(\tilde{u})\|_1 \leq 2\|u - \tilde{u}\|_\infty = \frac{2}{\tau}\|Q(s,\cdot) - \tilde{Q}(s,\cdot)\|_\infty$, which is (155). □

Thus, we can replace the $\arg\max$ in (28) by a softmax policy. This allows us to consider more verifiable assumptions and still derive the convergence. See our discussion in Section D.4.

It then suffices to verify the Lipschitz continuity of the worst-case kernel. We study it under three different distributional uncertainty sets.

D.1.2 Total variation balls

Consider the total variation ambiguity family:

$$\mathfrak{P}_{\mathrm{TV}}(s,a,\mu) \;:=\; \Big\{ p \in \Delta(\mathcal{S}) :\; \|p - \bar{p}_{s,a}(\mu)\|_1 \leq \varepsilon_{s,a} \Big\}, \qquad \varepsilon_{s,a} \geq 0. \tag{156}$$

Assume the center is Lipschitz:

$$\max_{(s,a)} \|\bar{p}_{s,a}(\mu) - \bar{p}_{s,a}(\tilde{\mu})\|_1 \leq L_{\bar{p}} \|\mu - \tilde{\mu}\|_1. \tag{157}$$

Any selection $p_\mu(\cdot \mid s, a) \in \mathfrak{P}_{\mathrm{TV}}(s,a,\mu)$ satisfies

$$\|p_\mu(\cdot \mid s, a) - p_{\tilde{\mu}}(\cdot \mid s, a)\|_1 \leq \|\bar{p}_{s,a}(\mu) - \bar{p}_{s,a}(\tilde{\mu})\|_1 + 2\varepsilon_{s,a} \leq L_{\bar{p}} \|\mu - \tilde{\mu}\|_1 + 2\varepsilon_{s,a}. \tag{158}$$

Consequently,

$$\max_s \|K_\mu(\cdot \mid s) - K_{\tilde{\mu}}(\cdot \mid s)\|_1 \leq (L_\pi + L_{\bar{p}}) \|\mu - \tilde{\mu}\|_1 + 2\bar{\varepsilon}, \qquad \bar{\varepsilon} := \max_{s,a} \varepsilon_{s,a}. \tag{159}$$

Plugging (159) into the recursion yields convergence to an $O(\bar{\varepsilon})$-neighborhood via Theorem 5 (interpret $2\bar{\varepsilon}$ as a deterministic per-iteration update error).

D.1.3 Wasserstein–1 balls

Consider the W_1 ambiguity family: fix a ground metric d on S and define

$$\mathfrak{P}_{\mathrm{W}}(s,a,\mu) \;:=\; \Big\{p \in \Delta(\mathcal{S}) :\; W_1\big(p, \bar{p}_{s,a}(\mu)\big) \le \rho_{s,a}(\mu)\Big\}, \qquad \rho_{s,a}(\mu) \ge 0. \tag{160}$$

Let

$$\underline{d} \;:=\; \min\{d(x,y) : x \neq y\} \in (0,\infty), \qquad \bar{\rho} \;:=\; \max_{s,a,\mu} \rho_{s,a}(\mu). \tag{161}$$

Lemma 16. *For all $p, q \in \Delta(\mathcal{S})$,*

$$\|p - q\|_1 \le \frac{2}{\underline{d}}\, W_1(p,q). \tag{162}$$

Proof. For any coupling $\gamma \in \Gamma(p,q)$, we have $d(X,Y) \ge \underline{d}\,\mathbf{1}\{X \neq Y\}$ almost surely, so $\mathbb{E}[d(X,Y)] \ge \underline{d}\,\mathbb{P}(X \neq Y) \ge \underline{d}\, d_{\mathrm{TV}}(p,q) = \underline{d}\,\|p-q\|_1/2$. Taking the infimum over γ yields (162). □

Assume the center is Lipschitz in W_1:

$$\max_{(s,a)} W_1\big(\bar{p}_{s,a}(\mu), \bar{p}_{s,a}(\tilde{\mu})\big) \le L_{\bar{p}}^{\mathrm{W}} \|\mu - \tilde{\mu}\|_1. \tag{163}$$

Then for any selections $p_\mu(\cdot \mid s,a) \in \mathfrak{P}_{\mathrm{W}}(s,a,\mu)$ and $p_{\tilde{\mu}}(\cdot \mid s,a) \in \mathfrak{P}_{\mathrm{W}}(s,a,\tilde{\mu})$,

$$\begin{aligned}
\|p_\mu(\cdot \mid s,a) - p_{\tilde{\mu}}(\cdot \mid s,a)\|_1 &\le \frac{2}{\underline{d}}\, W_1\big(p_\mu(\cdot \mid s,a), p_{\tilde{\mu}}(\cdot \mid s,a)\big) \\
&\le \frac{2}{\underline{d}}\Big(\rho_{s,a}(\mu) + W_1\big(\bar{p}_{s,a}(\mu), \bar{p}_{s,a}(\tilde{\mu})\big) + \rho_{s,a}(\tilde{\mu})\Big) \\
&\le \frac{2}{\underline{d}}\Big(L_{\bar{p}}^{\mathrm{W}} \|\mu - \tilde{\mu}\|_1 + 2\bar{\rho}\Big).
\end{aligned} \tag{164}$$

Therefore,

$$\max_s \|K_\mu(\cdot \mid s) - K_{\tilde{\mu}}(\cdot \mid s)\|_1 \le \Big(L_\pi + \frac{2}{\underline{d}} L_{\bar{p}}^{\mathrm{W}}\Big)\|\mu - \tilde{\mu}\|_1 + \frac{4\bar{\rho}}{\underline{d}}. \tag{165}$$

As in the TV case, (165) yields convergence to an $O(\bar{\rho})$-neighborhood via Theorem 5. Exact contraction (with no additive slack) typically requires additional structure (e.g., a unique stable optimizer selection or a regularized OT-based inner minimization).

D.2 Convergence

Lemma 17. *Let $\mu \in \Delta(\mathcal{S})$ and let (π_μ, p_μ) be any selections as in (28)–(29), i.e., $\operatorname{supp}\pi_\mu(\cdot|s) \subseteq \mathsf{D}(s,\mu)$ and $p_\mu(\cdot|s,a) \in \widehat{\mathfrak{P}}(s,a,\mu)$ for all (s,a). Then:*

(i) (robust optimality and worst-case attainment at μ)

$$V_\mu \;=\; \inf_{p \in \mathfrak{P}(\mu)} J_\mu(\pi_\mu, p) \;=\; J_\mu(\pi_\mu, p_\mu) \;=\; \langle \mu,\, v_\mu \rangle\,.$$

(ii) If in addition $\mu = F(\mu) = \mu K_\mu$, then (μ, π_μ, p_μ) is a stationary robust mean-field equilibrium in the sense of Definition 2.

Proof. Fix the population μ and apply Proposition 1 at μ: since π_μ is supported on the greedy sets $\mathsf{D}(\cdot,\mu)$ and $p_\mu(\cdot|s,a) \in \widehat{\mathfrak{P}}(s,a,\mu)$ pointwise, Proposition 1 gives, for every $s \in \mathcal{S}$,

$$v_\mu(s) \;=\; \inf_{p \in \mathfrak{P}(\mu)} J_\mu(s; \pi_\mu, p) \;=\; J_\mu(s; \pi_\mu, p_\mu), \qquad \text{and} \qquad \inf_{p \in \mathfrak{P}(\mu)} J_\mu(s; \pi', p) \;\le\; v_\mu(s) \;\; \forall \pi'. \tag{166}$$

Averaging the second identity in (166) against μ gives $J_\mu(\pi_\mu, p_\mu) = \sum_s \mu(s)\, J_\mu(s; \pi_\mu, p_\mu) = \langle \mu,\, v_\mu \rangle$.

Then, for any admissible stationary kernel $p \in \mathfrak{P}(\mu)$, the first identity in (166) gives $J_\mu(s; \pi_\mu, p) \ge v_\mu(s)$ for every s, hence $J_\mu(\pi_\mu, p) \ge \langle \mu,\, v_\mu \rangle$.

Therefore, the infimum over p of $J_\mu(\pi_\mu, p)$ equals $\langle \mu,\, v_\mu \rangle$ and is attained at p_μ. Finally, Proposition 2 gives $V_\mu = V_\mu(\mu) = \langle \mu,\, v_\mu \rangle$, completing (i); in particular Definition 2(i)–(ii) hold at the population μ.

For (ii), it remains only to check Definition 2(iii). The hypothesis $\mu = \mu K_\mu$ reads, coordinatewise,

$$\mu(s') \;=\; \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} \pi_\mu(a|s)\, p_\mu(s'|s,a) \qquad \forall s' \in \mathcal{S},$$

which is the consistency condition for the triple (μ, π_μ, p_μ). □

Lemma 18 (Contraction of the population operator)**.** *Under Assumption 5, the map $F(\mu) = \mu K_\mu$ is a contraction in $\|\cdot\|_1$:*

$$\|F(\mu) - F(\tilde\mu)\|_1 \le \rho\, \|\mu - \tilde\mu\|_1, \qquad \forall \mu, \tilde\mu \in \Delta(\mathcal{S}), \tag{167}$$

where $\rho = \rho_{\mathrm{mix}} + L_K < 1$.

Proof. Add and subtract $\mu K_{\tilde\mu}$:

$$\|F(\mu) - F(\tilde\mu)\|_1 = \|\mu K_\mu - \tilde\mu K_{\tilde\mu}\|_1 \le \|\mu K_\mu - \mu K_{\tilde\mu}\|_1 + \|\mu K_{\tilde\mu} - \tilde\mu K_{\tilde\mu}\|_1. \tag{168}$$

For the second term, first note that the Dobrushin contraction inequality implies

$$\|\nu K - \tilde\nu K\|_1 \le \alpha(K)\, \|\nu - \tilde\nu\|_1, \quad \forall \nu, \tilde\nu \in \Delta(\mathcal{S}). \tag{169}$$

Apply (169) and (34):

$$\|\mu K_{\tilde\mu} - \tilde\mu K_{\tilde\mu}\|_1 \le \alpha(K_{\tilde\mu})\|\mu - \tilde\mu\|_1 \le \rho_{\mathrm{mix}}\|\mu - \tilde\mu\|_1. \tag{170}$$

For the first term,

$$\|\mu K_\mu - \mu K_{\tilde\mu}\|_1 = \left\| \sum_{s\in S} \mu(s) \big(K_\mu(\cdot \mid s) - K_{\tilde\mu}(\cdot \mid s)\big) \right\|_1 \le \sum_{s\in S} \mu(s) \|K_\mu(\cdot \mid s) - K_{\tilde\mu}(\cdot \mid s)\|_1 \le \max_s \|K_\mu(\cdot \mid s) - K_{\tilde\mu}(\cdot \mid s)\|_1. \tag{171}$$

Now apply (35) to get $\|\mu K_\mu - \mu K_{\tilde\mu}\|_1 \le L_K \|\mu - \tilde\mu\|_1$. Combining the two bounds yields (167). □

Theorem 10. *Under Assumption 5, the map F has a unique fixed point $\mu^\star$. Moreover, the damped RB-Picard update*

$$\mu^{k+1} = (1-\alpha)\mu^k + \alpha F(\mu^k) \tag{172}$$

converges linearly to $\mu^\star$:

$$\|\mu^k - \mu^\star\|_1 \;\le\; \big(1 - \alpha(1-\rho)\big)^k\, \|\mu^0 - \mu^\star\|_1, \qquad \forall k \ge 0. \tag{173}$$

In particular, with $\alpha = 1$ (undamped iteration), $\|\mu^k - \mu^\star\|_1 \le \rho^k \|\mu^0 - \mu^\star\|_1$.

Proof. By Lemma 18, F is a contraction on the complete metric space $(\Delta(\mathcal{S}), \|\cdot\|_1)$, so Banach's fixed point theorem yields a unique fixed point $\mu^\star$ and linear convergence for $\alpha = 1$. For the damped update (172),

$$\|\mu^{k+1} - \mu^\star\|_1 \le (1-\alpha)\|\mu^k - \mu^\star\|_1 + \alpha\|F(\mu^k) - F(\mu^\star)\|_1 \le (1 - \alpha + \alpha\rho)\|\mu^k - \mu^\star\|_1 = \big(1 - \alpha(1-\rho)\big)\|\mu^k - \mu^\star\|_1, \tag{174}$$

and iterating gives (173). □

Theorem 11 (Stability under inexact updates)**.** *Assume Assumption 5 and that Algorithm 1 uses the inexact update $\mu^{k+1} = (1-\alpha)\mu^k + \alpha\widehat{F}(\mu^k)$ with* (36)*. Then, letting $q := 1 - \alpha(1-\rho) \in (0,1)$,*

$$\|\mu^k - \mu^\star\|_1 \;\le\; q^k \|\mu^0 - \mu^\star\|_1 \;+\; \alpha \sum_{j=0}^{k-1} q^{k-1-j} \varepsilon_j. \tag{175}$$

In particular, if $\sup_j \varepsilon_j \le \bar\varepsilon$, then $\limsup_{k\to\infty} \|\mu^k - \mu^\star\|_1 \le \bar\varepsilon/(1-\rho)$.

Proof. Write the recursion with the fixed point $\mu^\star = F(\mu^\star)$:

$$\mu^{k+1} - \mu^\star = (1-\alpha)(\mu^k - \mu^\star) + \alpha\big(\widehat{F}(\mu^k) - F(\mu^\star)\big). \tag{176}$$

Add and subtract $F(\mu^k)$ and use triangle inequality:

$$\|\mu^{k+1} - \mu^\star\|_1 \le (1-\alpha)\|\mu^k - \mu^\star\|_1 + \alpha\|F(\mu^k) - F(\mu^\star)\|_1 + \alpha\|\widehat{F}(\mu^k) - F(\mu^k)\|_1. \tag{177}$$

Apply Lemma 18 and (36):

$$\|\mu^{k+1} - \mu^\star\|_1 \le (1 - \alpha + \alpha\rho)\|\mu^k - \mu^\star\|_1 + \alpha\varepsilon_k = q\|\mu^k - \mu^\star\|_1 + \alpha\varepsilon_k. \tag{178}$$

Unrolling yields (175). The $\limsup$ statement follows by bounding the geometric series. □

D.3 Convergence under mixing setting

Assume $\alpha(K_\mu) \le \rho_{\mathrm{mix}} < 1$ for all μ so that each K_μ admits a unique stationary distribution. Define the stationary-distribution map $\Phi : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ by

$$\Phi(\mu) \;:=\; \text{the unique } \hat{\mu} \in \Delta(\mathcal{S}) \text{ such that } \hat{\mu} = \hat{\mu} K_\mu. \tag{179}$$

Theorem 12 (Contraction of the stationary-distribution map)**.** *Suppose* (34) *and* (35) *hold with* $\rho_{\mathrm{mix}} < 1$*. Then* Φ *is Lipschitz with*

$$\|\Phi(\mu) - \Phi(\tilde{\mu})\|_1 \le \frac{L_K}{1 - \rho_{\mathrm{mix}}} \|\mu - \tilde{\mu}\|_1. \tag{180}$$

Consequently, if $L_K < 1 - \rho_{\mathrm{mix}}$ *(equivalently* $\rho_{\mathrm{mix}} + L_K < 1$*), then* Φ *is a contraction.*

Proof. Let $\hat{\mu} = \Phi(\mu)$ and $\hat{\tilde{\mu}} = \Phi(\tilde{\mu})$. Then $\hat{\mu} = \hat{\mu} K_\mu$ and $\hat{\tilde{\mu}} = \hat{\tilde{\mu}} K_{\tilde{\mu}}$, so

$$\hat{\mu} - \hat{\tilde{\mu}} = \hat{\mu} K_\mu - \hat{\tilde{\mu}} K_{\tilde{\mu}} = (\hat{\mu} - \hat{\tilde{\mu}}) K_\mu + \hat{\tilde{\mu}} (K_\mu - K_{\tilde{\mu}}). \tag{181}$$

Taking $\|\cdot\|_1$ norms and applying (169) and (35) yields

$$\|\hat{\mu} - \hat{\tilde{\mu}}\|_1 \le \alpha(K_\mu) \|\hat{\mu} - \hat{\tilde{\mu}}\|_1 + \max_s \|K_\mu(\cdot \mid s) - K_{\tilde{\mu}}(\cdot \mid s)\|_1 \le \rho_{\mathrm{mix}} \|\hat{\mu} - \hat{\tilde{\mu}}\|_1 + L_K \|\mu - \tilde{\mu}\|_1. \tag{182}$$

Rearranging gives (180). □

D.4 Softmax robust best response and Soft Algorithm

Fix $\tau > 0$. For a given population distribution $\mu \in \Delta(\mathcal{S})$ and $v \in \mathbb{R}^S$, define the robust Q-operator

$$(Q_\mu v)(s,a) := \min_{P \in \mathfrak{P}(s,a,\mu)} \sum_{s' \in S} P(s') \big(r(s,a,s',\mu) + \gamma v(s') \big), \tag{183}$$

and the *soft* robust Bellman operator

$$(T_\mu^\tau v)(s) := \tau \log \sum_{a \in A} \exp\Big(\frac{(Q_\mu v)(s,a)}{\tau} \Big). \tag{184}$$

Lemma 19. *For each fixed* μ*, the map* T_μ^τ *is a contraction on* $(\mathbb{R}^S, \|\cdot\|_\infty)$ *with modulus* γ*. Hence there exists a unique* v_μ^τ *such that* $v_\mu^\tau = T_\mu^\tau v_\mu^\tau$*.*

Proof. For any s, a and any v, w, one has $|(Q_\mu v)(s,a) - (Q_\mu w)(s,a)| \le \gamma \|v - w\|_\infty$. Also, $\mathrm{LSE}_\tau(x) := \tau \log \sum_a e^{x_a/\tau}$ satisfies $|\mathrm{LSE}_\tau(x) - \mathrm{LSE}_\tau(y)| \le \|x - y\|_\infty$. Combining yields $\|T_\mu^\tau v - T_\mu^\tau w\|_\infty \le \gamma \|v - w\|_\infty$. □

Define the soft robust Q-function $Q_\mu^\tau(s,a) := (Q_\mu v_\mu^\tau)(s,a)$ and the SoftMax policy

$$\pi_\mu^\tau(a \mid s) := \frac{\exp(Q_\mu^\tau(s,a)/\tau)}{\sum_{b \in A} \exp(Q_\mu^\tau(s,b)/\tau)}. \tag{185}$$

Choose a worst-case kernel selector

$$p_\mu^\tau(\cdot \mid s,a) \in \arg\min_{P \in \mathfrak{P}(s,a,\mu)} \sum_{s' \in S} P(s') \big(r(s,a,s',\mu) + \gamma v_\mu^\tau(s') \big). \tag{186}$$

Let $K_\mu^\tau(s' \mid s) := \sum_a \pi_\mu^\tau(a \mid s) p_\mu^\tau(s' \mid s,a)$ and define the population operator $F^\tau(\mu) := \mu K_\mu^\tau$.

Assumption 8 (Soft contractivity for F^τ)**.** *There exist* $\rho_{\mathrm{mix}} \in [0,1)$*,* $L_Q^\tau \ge 0$*, and* $L_p^\tau \ge 0$ *such that for all* $\mu, \tilde{\mu}$*: (i)* $\alpha(K_\mu^\tau) \le \rho_{\mathrm{mix}}$*; (ii)* $\max_{s,a} |Q_\mu^\tau(s,a) - Q_{\tilde{\mu}}^\tau(s,a)| \le L_Q^\tau \|\mu - \tilde{\mu}\|_1$*; (iii)* $\max_{s,a} \|p_\mu^\tau(\cdot \mid s,a) - p_{\tilde{\mu}}^\tau(\cdot \mid s,a)\|_1 \le L_p^\tau \|\mu - \tilde{\mu}\|_1$*. Assume* $\rho^\tau := \rho_{\mathrm{mix}} + L_p^\tau + \frac{2 L_Q^\tau}{\tau} < 1$.

Theorem 13 (Linear convergence)**.** *Under Assumption 8,* F^τ *has a unique fixed point* $\mu^{\tau,\star}$*. Moreover, the update* $\mu^{k+1} = (1 - \alpha)\mu^k + \alpha F^\tau(\mu^k)$ *satisfies*

$$\|\mu^k - \mu^{\tau,\star}\|_1 \le (1 - \alpha(1 - \rho^\tau))^k \|\mu^0 - \mu^{\tau,\star}\|_1. \tag{187}$$

Algorithm 2 Soft RB-Iteration

Require: initial $\mu^0 \in \Delta(\mathcal{S})$; temperature $\tau > 0$; stepsize $\alpha \in (0, 1]$; tolerances $\varepsilon_{\mathrm{DP}}, \varepsilon_\mu$.

for $k = 0, 1, 2, \ldots$ **do**
  (Soft robust DP at μ^k) iterate $v^{(m+1)} \leftarrow T^\tau_{\mu^k} v^{(m)}$ until $\|v^{(m+1)} - v^{(m)}\|_\infty \le \varepsilon_{\mathrm{DP}}$; set $v^\tau_{\mu^k} \leftarrow v^{(m+1)}$.
  (SoftMax policy) set $Q^k(s,a) := (Q_{\mu^k} v^\tau_{\mu^k})(s,a)$ and $\pi^k(a \mid s) \propto \exp(Q^k(s,a)/\tau)$.
  (Worst-case kernel) for each (s,a) choose $p^k(\cdot \mid s,a) \in \arg\min_{P \in \mathfrak{P}(s,a,\mu^k)} \sum_{s'} P(s') \big(r + \gamma v^\tau_{\mu^k}\big)$.
  form $K^k(s' \mid s) = \sum_a \pi^k(a \mid s) p^k(s' \mid s, a)$ and $\tilde{\mu}^{k+1} \leftarrow \mu^k K^k$.
  (Damped update) $\mu^{k+1} \leftarrow (1-\alpha)\mu^k + \alpha\, \tilde{\mu}^{k+1}$.
  if $\|\mu^{k+1} - \mu^k\|_1 \le \varepsilon_\mu$ **then**
    break
  end if
end for
return (μ^{k+1}, π^k, p^k).

Proof. Combine: (a) Dobrushin contraction in the initial measure with factor ρ_{mix}, (b) the kernel Lipschitz bound $\max_s \|K^\tau_\mu(\cdot \mid s) - K^\tau_{\tilde{\mu}}(\cdot \mid s)\|_1 \le L^\tau_p \|\mu - \tilde{\mu}\|_1 + \frac{2L^\tau_Q}{\tau}\|\mu - \tilde{\mu}\|_1$ (using SoftMax Lipschitz), to show $\|F^\tau(\mu) - F^\tau(\tilde{\mu})\|_1 \le \rho^\tau \|\mu - \tilde{\mu}\|_1$. Then apply Banach fixed point theorem and the standard damped contraction recursion. □

We moreover quantify the closeness of the soft limit and the original one.

Proposition 7 (Bias of the soft equilibrium)**.** *Let $\tau > 0$, let $\mu^{\tau,\star} = F^\tau(\mu^{\tau,\star})$ with associated (π^τ, p^τ), and set $\varepsilon(\tau) := \dfrac{\tau \log|\mathcal{A}|}{1-\gamma}$. Write $\mu := \mu^{\tau,\star}$. Then:*

(i) $v_\mu \le v^\tau_\mu \le v_\mu + \varepsilon(\tau)$ *pointwise;*

(ii) $v^\tau_\mu - \varepsilon(\tau) \le u^{\pi^\tau}_\mu \le v^\tau_\mu$, *hence* $\inf_p J_\mu(s; \pi^\tau, p) \ge v_\mu(s) - \varepsilon(\tau)$ *for all s;*

(iii) $u^{\pi^\tau}_\mu \le J_\mu(\cdot; \pi^\tau, p^\tau) \le u^{\pi^\tau}_\mu + \dfrac{2\gamma\, \varepsilon(\tau)}{1-\gamma}$ *pointwise.*

Consequently $(\mu^{\tau,\star}, \pi^\tau, p^\tau)$ is an $\big(\varepsilon(\tau), \frac{2\gamma}{1-\gamma}\varepsilon(\tau)\big)$-robust MFE, and τ trades equilibrium accuracy against the verifiability ($L_\pi = 2L^\tau_Q/\tau$).

Proof. Throughout, $\varepsilon_0 := \tau \log|\mathcal{A}|$, so $\varepsilon(\tau) = \varepsilon_0/(1-\gamma)$. We use the elementary perturbation fact: if S is a monotone γ-contraction with $S(v + c\mathbf{1}) = Sv + \gamma c\mathbf{1}$, and $Sw \ge w - \varepsilon_0\mathbf{1}$ pointwise, then its fixed point u satisfies $u \ge w - \varepsilon(\tau)\mathbf{1}$[6]. Symmetrically $Sw \le w + \varepsilon_0\mathbf{1}$ implies $u \le w + \varepsilon(\tau)\mathbf{1}$, and $S_1 \le S_2$ pointwise (both monotone contractions) implies their fixed points are ordered.

(i) $\max_a x_a \le \mathrm{LSE}_\tau(x) \le \max_a x_a + \tau \log|\mathcal{A}|$ gives $\mathcal{T}_\mu v \le \mathcal{T}^\tau_\mu v \le \mathcal{T}_\mu v + \varepsilon_0\mathbf{1}$ for every v; the ordering of fixed points gives $v_\mu \le v^\tau_\mu$, and applying the perturbation fact to $S = \mathcal{T}_\mu$, $w = v^\tau_\mu$ (note $\mathcal{T}_\mu v^\tau_\mu \ge \mathcal{T}^\tau_\mu v^\tau_\mu - \varepsilon_0\mathbf{1} = v^\tau_\mu - \varepsilon_0\mathbf{1}$) gives $v_\mu \ge v^\tau_\mu - \varepsilon(\tau)$.

(ii) The Fenchel identity $\mathrm{LSE}_\tau(x) = \sum_a \pi^\tau_a x_a + \tau H(\pi^\tau)$ with $\pi^\tau = \mathrm{softmax}(x/\tau)$ and $H \le \log|\mathcal{A}|$ gives, at $x = Q^\tau_\mu(s,\cdot)$, $(\mathcal{T}_\mu^{\pi^\tau} v^\tau_\mu)(s) = v^\tau_\mu(s) - \tau H(\pi^\tau(\cdot|s)) \in [v^\tau_\mu(s) - \varepsilon_0,\, v^\tau_\mu(s)]$. Apply the perturbation fact (both directions) to $S = \mathcal{T}_\mu^{\pi^\tau}$, $w = v^\tau_\mu$: its fixed point $u^{\pi^\tau}_\mu$ satisfies $v^\tau_\mu - \varepsilon(\tau) \le u^{\pi^\tau}_\mu \le v^\tau_\mu$.

(iii) The lower bound is $J_\mu(\cdot; \pi^\tau, p^\tau) \ge \inf_p J_\mu(\cdot; \pi^\tau, p) = u^{\pi^\tau}_\mu$. For the upper bound, write $u := u^{\pi^\tau}_\mu$, $c_v := r + \gamma v$. For each (s,a), with p^τ the minimizer of $\langle \cdot, c_{v^\tau_\mu} \rangle$ over $\mathfrak{P}(s,a,\mu)$,

$$\langle p^\tau, c_u \rangle \le \langle p^\tau, c_{v^\tau_\mu} \rangle + \gamma \|u - v^\tau_\mu\|_\infty = \min_P \langle P, c_{v^\tau_\mu} \rangle + \gamma \|u - v^\tau_\mu\|_\infty \le \min_P \langle P, c_u \rangle + 2\gamma \|u - v^\tau_\mu\|_\infty.$$

Since $\|u - v^\tau_\mu\|_\infty \le \varepsilon(\tau)$ by (ii), averaging over $\pi^\tau(\cdot|s)$ yields $\mathcal{T}_\mu^{\pi^\tau, p^\tau} u \le \mathcal{T}_\mu^{\pi^\tau} u + 2\gamma\varepsilon(\tau)\mathbf{1} = u + 2\gamma\varepsilon(\tau)\mathbf{1}$. The perturbation fact applied to $S = \mathcal{T}_\mu^{\pi^\tau, p^\tau}$, whose fixed point is $J_\mu(\cdot; \pi^\tau, p^\tau)$, gives the claim with constant $2\gamma\varepsilon(\tau)/(1-\gamma)$. □

[6] The results are from induction: $S^{n+1} w \ge S(w - \varepsilon_0 \sum_{i<n} \gamma^i \mathbf{1}) = Sw - \varepsilon_0 \sum_{1 \le i \le n} \gamma^i \mathbf{1} \ge w - \varepsilon_0 \sum_{i \le n} \gamma^i \mathbf{1}$; let $n \to \infty$.