MARKOV-NASH EQUILIBRIA IN MEAN-FIELD GAMES UNDER MODEL UNCERTAINTY

JOHANNES LANGNER, ARIEL NEUFELD, AND KYUNGHYUN PARK

Abstract. We propose and analyze a framework for mean-field Markov games under model uncertainty. In this framework, a state-measure flow describing the collective behavior of a population affects the given reward function as well as the *unknown* transition kernel of the representative agent. The agent's objective is to choose an optimal Markov policy in order to maximize her worst-case expected reward, where worst-case refers to the most adverse scenario among all transition kernels considered to be feasible to describe the unknown true law of the environment. We prove the existence of a mean-field equilibrium under model uncertainty, where the agent chooses the optimal policy that maximizes the worst-case expected reward, and the state-measure flow aligns with the agent's state distribution under the optimal policy and the worst-case transition kernel. Moreover, we prove that for suitable multi-agent Markov games under model uncertainty the optimal policy from the mean-field equilibrium forms an approximate Markov-Nash equilibrium whenever the number of agents is large enough.

Contents

Date: November 11, 2024.

Key words: mean-field games, Nash equilibrium, Markov decision processes, model uncertainty, distributionally robust optimization.

MSC2020 Subject Classification: 91A16, 91A10, 90C40, 90C17.

Funding: J. Langner is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 471178162. A. Neufeld gratefully acknowledges support by the MOE AcRF Tier 2 Grant MOE-T2EP20222-0013. K. Park acknowledges support of the Presidential Postdoctoral Fellowship of the Nanyang Technological University.

1. Introduction

Mean-field games introduced by [34, 40] analyze decision-making and interactions of strategic agents within populations. Under the assumption that all agents of a population have the same transition probabilities and reward function and that their interactions only depend on the empirical distribution of all agents, one can simplify the model by approximating the finite agent game by a suitable mean-field game. This framework has led to a wide range of applications, including in finance and economics (e.g., [11,13,14,39,42]), crowd motion dynamics (e.g., [32,36]), and epidemiology (e.g., [3, 19]).

As a prominent discrete-time mean-field games model, consider a mean-field Markov game denoted by (S, A, μ^o, p, r): Let (S, A) be state and action spaces and denote by $\mathcal{P}(S)$ and $\mathcal{P}(A)$ the set of probability measures on S and A, respectively. Furthermore, let $\mu^o \in \mathcal{P}(S)$ be an initial population distribution, $p : S\times A\times\mathcal{P}(S) \mapsto \mathcal{P}(S)$ be a transition kernel, and $r : S\times A\times S\times\mathcal{P}(S) \mapsto \mathbb{R}$ be a one-step reward function. Assume that a representative agent aims to maximize her total expected reward until the terminal time T by choosing a Markov policy $\pi_{0:T} = (\pi_0, \ldots, \pi_{T-1})$ (i.e., a sequence of stochastic kernels $\pi_t : S \mapsto \mathcal{P}(A)$, $t = 0, \ldots, T-1$). Given a population measure flow $\mu_{0:T} = (\mu_0, \ldots, \mu_{T-1})$ with $\mu_0 = \mu^o$ (i.e., a sequence of $\mu_t \in \mathcal{P}(S)$, $t = 0, \ldots, T-1$), the central objective the agent faces is to solve the following Markov decision problem

$$\sup_{\pi_{0:T}} \mathbb{E}^{\mathbb{P}}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t)\right], \tag{1.1}$$

where for given $\pi_{0:T}$, $\mathbb{P}$ is the probability measure (that depends on $\mu_{0:T}, \pi_{0:T}$, and p) under which the agent's state and action configurations evolve as follows: for every $t = 0, \ldots, T-1$

$$s_0 \sim \mu_0(\cdot), \quad a_t \sim \pi_t(\cdot|s_t), \quad s_{t+1} \sim p(\cdot|s_t, a_t, \mu_t). \tag{1.2}$$

In this setting, a mean-field equilibrium consists of a Markov policy and a measure flow $(\mu^*_{0:T}, \pi^*_{0:T})$ satisfying that $\pi^*_{0:T}$ is a maximizer of (1.1) given $\mu^*_{0:T}$, and $\mu^*_{0:T}$ is consistent with the state distribution of the agent acting optimally via $\pi^*_{0:T}$, i.e., $\mu^*_0 = \mu^o$ and for $t = 0, \ldots, T-1$

$$\mu^*_{t+1}(ds_{t+1}) = \int_{S\times A} p(ds_{t+1}|s_t, a_t, \mu^*_t)\pi^*_t(da_t|s_t)\mu^*_t(ds_t). \tag{1.3}$$

In most cases, a mean-field equilibrium attains an approximate Nash equilibrium for an analogous game with a finite number of agents, known as the so-called Nash certainty equivalence principle [6,10,12,34]. We refer to [20,21,24,26,43,51,52] for a few articles studying discrete-time mean-field games similar to the setting (S, A, μ^o, p, r) described above.

Mean-field games commonly involve a significant assumption that the model environment represented by the transition kernel p in the above model (S, A, μ^o, p, r) is perfectly known to all agents. However, when implemented in practice, the specifics of the model environment are a priori unclear. While some estimation techniques can approximate a ground truth on, e.g., the transition kernel closely, in many cases there exists a margin of misspecification. This might result in an equilibrium that is not consistent with the behavior of large populations in real situations.

As a remedy to *model uncertainty*, a number of researchers in various fields have adopted the so-called worst-case (or robust) approach introduced by [16, 18, 22, 25]. Here, worst-case refers to considering the most adverse scenario among all probabilities deemed as feasible to describe the unknown law characterizing the environment. The aim of this article is to propose and analyze a framework for mean-field Markov games under model uncertainty, which can be considered as a robust analog of (S, A, μ^o, p, r) described in (1.1)-(1.3).

To that end, let us describe our mean-field Markov game under model uncertainty, which we denote by $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$: Fix $T \in \mathbb{N}$ and let (S, A, μ^o, r) be the same as the ones given in (S, A, μ^o, p, r) described above. Furthermore, let $\mathfrak{P}_{0:T}$ be a sequence of *set-valued* maps given for every $t = 0, \dots, T-1$ by

$$\mathfrak{P}_t : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \twoheadrightarrow \mathfrak{P}_t(s_t, a_t, \mu_t) \subseteq \mathcal{P}(S). \tag{1.4}$$

Then given $(\mu_{0:T}, \pi_{0:T})$, denote by $\mathcal{Q}(\mu_{0:T}, \pi_{0:T})$ the set of all probability measures $\mathbb{P}$ under which there exists a sequence of transition kernels $p_{0:T} = (p_0, \dots, p_{T-1})$ satisfying that for every $t = 0, \dots, T-1$ and every $(s_t, a_t, \mu_t) \in S \times A \times \mathcal{P}(S)$

$$p_t(\cdot|s_t, a_t, \mu_t) \in \mathfrak{P}_t(s_t, a_t, \mu_t), \tag{1.5}$$

and the agent's state and action configurations evolve as follows: for every $t = 0, \dots, T-1$

$$s_0 \sim \mu_0(\cdot), \quad a_t \sim \pi_t(\cdot|s_t), \quad s_{t+1} \sim p_t(\cdot|s_t, a_t, \mu_t). \tag{1.6}$$

In other words, instead of fixing a transition kernel $p : S \times A \times \mathcal{P}(S) \mapsto \mathcal{P}(S)$, we consider a set valued map $\mathfrak{P}_t : S \times A \times \mathcal{P}(S) \twoheadrightarrow \mathcal{P}(S)$ where given $(s_t, a_t, \mu_t) \in S \times A \times \mathcal{P}(S)$, each element of the set $\mathfrak{P}_t(s_t, a_t, \mu_t)$ is considered as a candidate probability measure on S derived from the true but unknown transition kernel. This setting is inspired by [48, 49] which analyzed Markov decision problems under model uncertainty (but without a mean-field measure flow).

Now, given $\mu_{0:T}$ with $\mu_0 = \mu^o$, the central objective an agent faces under model uncertainty is to solve the following robust (or worst-case) optimization problem

$$V(\mu_{0:T}) = \sup_{\pi_{0:T}} \inf_{\mathbb{P} \in \mathcal{Q}(\mu_{0:T}, \pi_{0:T})} \mathbb{E}^{\mathbb{P}} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t) \right]. \tag{1.7}$$

We note that the set-valued maps $\mathfrak{P}_{0:T}$ given in (1.4) induce distributional uncertainty represented by the set $\mathcal{Q}(\mu_{0:T}, \pi_{0:T})$, and (1.7) and (1.1) coincide when $\mathfrak{P}_{0:T}$ are singleton-valued.

In this setting, we say

$$(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T}) \tag{1.8}$$

is a mean-field equilibrium of $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ (see Definition 2.3) if the Markov policy $\pi^*_{0:T}$ is optimal to the robust optimization problem $V(\mu^*_{0:T})$, the transition kernel $p^*_{0:T}$ corresponds to the worst-case kernel of $V(\mu^*_{0:T})$ under $(\mu^*_{0:T}, \pi^*_{0:T})$, and the state-measure flow $\mu^*_{0:T}$ aligns with the agent's state distribution under $(\pi^*_{0:T}, p^*_{0:T})$, i.e., $\mu^*_0 = \mu^o$ and for every $t = 0, \dots, T-2$,

$$\mu^*_{t+1}(ds_{t+1}) = \int_{S \times A} p^*_t(ds_{t+1}|s_t, a_t, \mu^*_t) \pi^*_t(da_t|s_t) \mu^*_t(ds_t). \tag{1.9}$$

The main contribution of this paper is twofold:

- In Theorem 3.10, we prove the existence of a mean-field equilibrium $(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T})$ of the mean-field Markov game $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ described in (1.4)-(1.9).
- We show in Theorem 3.19 that the optimal Markov policy $\pi^*_{0:T}$ from the mean-field equilibrium of $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ forms an approximate Markov-Nash equilibrium (see Definition 2.6) of a multi-agent Markov game under model uncertainty in the sense that the policy $\pi^*_{0:T}$ is (almost) a maximizer for the worst-case objectives of all agents in the multi-agent Markov game (see (2.6)) whenever the number of agents is large enough.

As an example, in Section 4, we apply our mean-field Markov game $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ to crowd motion dynamics under model uncertainty. In this context, the set valued maps given in (1.4) are formulated by a Wasserstein-ball around a reference transition kernel (see Definition 4.1), which aligns with our conditions imposed on the set-valued maps in order to obtain our main results.

Moreover, we compute a mean-field equilibrium of the crowd motion dynamics by our iterative scheme (see Algorithm 1).

Related literature. Classic mean-field games (i.e., without uncertainty) are described both in continuous-time (see, e.g., [3,17,27,33,34,37–39,54]) and in discrete-time (see, e.g., [1,8,20,21,24, 26,43,46,50–52]); we refer to [6,12,28,41] for survey papers including both settings. We also refer to [4, 15, 23, 29, 30, 47] for mean-field control problems in a Markov decision process framework (which corresponds to cooperative models).

In continuous-time settings, several articles have explored mean-field games under distributional or parametric uncertainty (see., e.g., [5,31,44,45]). Notably, our notion of a mean-field equilibrium under model uncertainty (described in (1.8)–(1.9); see also Definition 2.3 and Theorem 3.10) aligns with those found in continuous-time frameworks (see, e.g., [5, Proposition 3], [31, Theorem 3.2]), where our robust optimization problem (1.7) corresponds in their papers to a forward backward system consisting of a Hamilton–Jacobi–Bellman–Isaacs equation, whereas our measure flow (1.9) corresponds in their papers to a Fokker-Planck equation under the associated worst-case measure or parameter. Moreover, [45, Theorem 6] establishes an approximate Nash equilibrium under model uncertainty, which is consistent with ours given in Theorem 3.19. To the best of our knowledge, however, there are no known results on mean-field games under model uncertainty in a discrete-time setting or within the framework of Markov decision processes.

While certain proof techniques in our paper bear similarities to [51,52] which consider mean-field Markov games in a discrete-time setting *but without model uncertainty*, the consideration of model uncertainty introduces significant distinctions. Specifically, due to the set-valued maps $\mathfrak{P}_{0:T}$ given in (1.4), we cannot directly apply certain existing arguments (including the dynamic programming principle and the fixed point approach). Instead, we establish a robust (i.e., max-min) version of the dynamic programming principle, which constitutes a variant of [49]. We then propose and study a robust analog of the fixed point approach based on the work of [35]. Moreover, we establish the dynamic programming principle for the multi-agent Markov game under model uncertainty and characterize the worst-case measures appearing in both the mean-field and multi-agent Markov games to establish the existence of an approximate Markov-Nash equilibrium.

2. Model description

2.1. **Notation and preliminaries.**

Throughout this article we work with Borel spaces. If X is such a space, we denote by $\mathcal{B}_X$ its Borel σ-field and $\mathcal{P}(X)$ the set of all probability measures on X implicitly assumed to be equipped with the topology induced by the weak convergence, i.e., for any $\mathbb{P} \in \mathcal{P}(X)$ and any $(\mathbb{P}^n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(X)$, we have

$$\mathbb{P}^n \rightharpoonup \mathbb{P} \text{ as } n \to \infty \;\Leftrightarrow\; \lim_{n\to\infty} \int_X f(\omega)\mathbb{P}^n(d\omega) = \int_X f(\omega)\mathbb{P}(d\omega) \text{ for any } f \in C_b(X;\mathbb{R}), \tag{2.1}$$

where $C_b(X;\mathbb{R})$ is the set of all continuous and bounded functions from X to $\mathbb{R}$.

If X is compact, the weak topology given in (2.1) is equivalent to the topology induced by the 1-Wasserstein distance $d_{W_1}(\cdot,\cdot)$ which we recall to be the following: For any $\mu,\nu \in \mathcal{P}(X)$, denote by $\mathrm{Cpl}(\mu,\nu) \subseteq \mathcal{P}(X \times X)$ the subset of all probability measures on $X \times X$ with first marginal μ and second marginal ν. Then the 1-Wasserstein distance between μ and ν is defined by

$$d_{W_1}(\mu,\nu) := \inf_{\gamma\in\mathrm{Cpl}(\mu,\nu)} \int_{X\times X} |x-y|\gamma(dx,dy),$$

where $|\cdot|$ is the Euclidean norm.

In particular, if we further assume that X is a finite subset in a Euclidean space and denote by $n(X)$ its cardinality, then $\mathcal{P}(X)$ can be identified with a simplex in $\mathbb{R}^{n(X)}$, i.e., $\mu \in \mathcal{P}(X)$ can be treated as an $n(X)$-dimensional vector $(w_1^\mu, \dots, w_{n(X)}^\mu) \in \mathbb{R}^{n(X)}$ with nonnegative coordinates $(w_i^\mu)_{i=1,\dots,n(X)}$ which sum up to one.

For each $t \in \mathbb{N}$, we use the abbreviation $X^t := X \times \cdots \times X$ for the t-times Cartesian product of the set X, where we endow X^t with the corresponding product topology. In analogy, we use $(\mathcal{P}(X))^t$ for the corresponding product of $\mathcal{P}(X)$. Given a sequence of probability measures $(\mathbb{P}_s, \dots, \mathbb{P}_{t-1}) \in (\mathcal{P}(X))^{s-t}$ and $0 \leq s < t$, we use the following abbreviation $\mathbb{P}_{s:t} := (\mathbb{P}_s, \dots, \mathbb{P}_{t-1})$. The same convention applies to a sequence of other quantities.

2.2. **Mean-field Markov games under model uncertainty.** We specify what we mean by mean-field Markov games under model uncertainty. Let us consider a representative agent who, at each time t, observes a state s_t and takes an action a_t, whereas a probability measures μ_t describes the overall population distribution at time t.

Definition 2.1 (Mean-field Markov game)**.** Fix a time horizon $T \in \mathbb{N}$. A mean-field Markov game under model uncertainty, say $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$, comprises the following:

(i) $(S, \mathcal{B}_S)$ and $(A, \mathcal{B}_A)$ are Borel spaces for the state and action spaces, respectively.
(ii) $\mu^o \in \mathcal{P}(S)$ is a given initial distribution for the initial state, which we denote by s_0.
(iii) For every $t = 0, \dots, T-1$, $\mathfrak{P}_t : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \twoheadrightarrow \mathfrak{P}_t(s_t, a_t, \mu_t) \subseteq \mathcal{P}(S)$ is a correspondence (i.e., a set-valued map) at time t, inducing distributional uncertainty in the next-state configuration.
(iv) $r : S \times A \times S \times \mathcal{P}(S) \mapsto \mathbb{R}$ is a one-step Borel-measurable reward function.

We proceed to describe the set of policies and the set of uncertain probability measures.

Definition 2.2. Let $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ be given in Definition 2.1.

(i) Define by Π the set of all sequences of Markov policies $\pi_{0:T}$ such that for $t = 0, \dots, T-1$, $\pi_t : S \ni s_t \mapsto \pi_t(\cdot|s_t) \in \mathcal{P}(A)$ is a so-called Markov kernel.
(ii) Given $(\mu_{0:T}, \pi_{0:T}) \in (\mathcal{P}(S))^T \times \Pi$ satisfying $\mu_0 = \mu^o$, we define by $\mathcal{Q}(\mu_{0:T}, \pi_{0:T}) \subseteq \mathcal{P}(S \times (S \times A)^T)$ the subset of all probability measures $\mathbb{P} := \mu_0 \otimes \mathbb{P}_{(\mu_0,\pi_0,p_0)} \otimes \cdots \otimes \mathbb{P}_{(\mu_{T-1},\pi_{T-1},p_{T-1})}$ such that[1] for every $t = 0, \dots, T-1$,

$$\mathbb{P}_{(\mu_t,\pi_t,p_t)} : S \ni s_t \mapsto \mathbb{P}_{(\mu_t,\pi_t,p_t)}(ds_{t+1}, da_t|s_t) := p_t(ds_{t+1}|s_t, a_t, \mu_t)\pi_t(da_t|s_t)$$

is a stochastic kernel[2] on $S \times A$ given S, and $p_t : S \times A \times \mathcal{P}(S) \mapsto \mathcal{P}(S)$ is a stochastic kernel satisfying that for every $(s_t, a_t, \mu_t) \in S \times A \times \mathcal{P}(S)$,

$$p_t(ds_{t+1}|s_t, a_t, \mu_t) \in \mathfrak{P}_t(s_t, a_t, \mu_t).$$

Denote by $V : (\mathcal{P}(S))^T \ni \mu_{0:T} \mapsto V(\mu_{0:T}) \in \mathbb{R}$ the robust optimization problem defined by

$$V(\mu_{0:T}) := \sup_{\pi_{0:T} \in \Pi} J(\mu_{0:T}, \pi_{0:T}), \tag{2.2}$$

[1] For every $t = 0, \dots, T-1$, $\mu_0 \otimes \mathbb{P}_{(\mu_0,\pi_0,p_0)} \otimes \cdots \otimes \mathbb{P}_{(\mu_t,\pi_t,p_t)}$ denotes an element in $\mathcal{P}(S \times (S \times A)^{t+1})$ satisfying that for every $B \in \mathcal{B}_{S \times (S \times A)^{t+1}}$,

$$\begin{aligned}&\mu_0 \otimes \mathbb{P}_{(\mu_0,\pi_0,p_0)} \otimes \cdots \otimes \mathbb{P}_{(\mu_t,\pi_t,p_t)}(B)\\ &:= \int_S \int_{S \times A} \cdots \int_{S \times A} \mathbf{1}_{\{(s_0,s_1,a_0,\dots,s_{t+1},a_t) \in B\}}\, \mathbb{P}_{(\mu_t,\pi_t,p_t)}(ds_{t+1}, da_t|s_t) \cdots \mathbb{P}_{(\mu_0,\pi_0,p_0)}(ds_1, da_0|s_0)\mu_0(ds_0).\end{aligned}$$

[2] Throughout the paper, a stochastic kernel p on X_2 given X_1, for some Borel spaces X_1 and X_2, is defined as a Borel-measurable mapping from X_1 to $\mathcal{P}(X_2)$.

where the worst-case objective $J : (\mathcal{P}(S))^T \times \Pi \ni (\mu_{0:T}, \pi_{0:T}) \mapsto J(\mu_{0:T}, \pi_{0:T}) \in \mathbb{R}$ is given by

$$J(\mu_{0:T}, \pi_{0:T}) := \inf_{\mathbb{P} \in \mathcal{Q}(\mu_{0:T}, \pi_{0:T})} \mathbb{E}^{\mathbb{P}} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t) \right]. \tag{2.3}$$

We now introduce what we refer to as a mean field equilibrium under model uncertainty.

Definition 2.3 (Mean-field equilibrium)**.** We call $(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T})$ a mean-field equilibrium of the mean-field Markov game $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ (see Definition 2.1) if the following conditions hold:

(i) $(\pi^*_{0:T}, p^*_{0:T})$ are optimal for $V(\mu^*_{0:T})$, i.e., $\pi^*_{0:T}$ is the optimal Markov policy of $V(\mu^*_{0:T})$ and $p^*_{0:T}$ is the worst-case transition kernel of $V(\mu^*_{0:T})$ under $(\mu^*_{0:T}, \pi^*_{0:T})$, i.e.,

$$\begin{aligned} V(\mu^*_{0:T}) = J(\mu^*_{0:T}, \pi^*_{0:T}) &= \sup_{\pi_{0:T} \in \Pi} \mathbb{E}^{\mathbb{P}(\mu^*_{0:T}, \pi_{0:T}, p^*_{0:T})} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu^*_t) \right] \\ &= \mathbb{E}^{\mathbb{P}^*} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu^*_t) \right], \end{aligned}$$

where for every $\pi_{0:T} \in \Pi$,

$$\mathbb{P}(\mu^*_{0:T}, \pi_{0:T}, p^*_{0:T}) := \mu^*_0 \otimes \mathbb{P}_{(\mu^*_0, \pi_0, p^*_0)} \otimes \cdots \otimes \mathbb{P}_{(\mu^*_{T-1}, \pi_{T-1}, p^*_{T-1})} \in \mathcal{Q}(\mu^*_{0:T}, \pi_{0:T}), \tag{2.4}$$

and $\mathbb{P}^* := \mathbb{P}(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T}) \in \mathcal{Q}(\mu^*_{0:T}, \pi^*_{0:T})$ (see Definition 2.2).

(ii) $\mu^*_{0:T}$ satisfies that $\mu^*_0(\cdot) = \mu^o(\cdot)$ and for every $t = 0, \dots, T-2$,

$$\mu^*_{t+1}(\cdot) = \int_{S \times A} p^*_t(\cdot | s_t, a_t, \mu^*_t) \pi^*_t(da_t | s_t) \mu^*_t(ds_t).$$

2.3. **Multi-agent Markov games under model uncertainty.** We aim to obtain approximate Markov-Nash equilibria under model uncertainty by using mean-field equilbria under model uncertainty. To that end, in this section, we introduce the framework for multi-agent Markov games under model uncertainty and the notion of their Markov-Nash equilibria.

Let $N \in \mathbb{N}$ be the number of agents and, as before, S and A be the state and action spaces, respectively. For $i = 1, \dots, N$, denote by $s^i_t \in S$ and $a^i_t \in A$ the state and action configurations of the agent i at time t, respectively. Then we set

$$\overline{s}^N_t := (s^1_t, \dots, s^N_t) \in S^N, \qquad \overline{a}^N_t := (a^1_t, \dots, a^N_t) \in A^N$$

to be the state and action configurations of all N agents at time t, respectively, and denote by

$$e^N(\overline{s}^N_t) := \frac{1}{N} \sum_{i=1}^N \delta_{s^i_t} \in \mathcal{P}(S) \tag{2.5}$$

the empirical distribution of $\overline{s}^N_t$, where $\delta_s \in \mathcal{P}(S)$ denotes the Dirac measure at $s \in S$.

Definition 2.4 (Multi-agent Makov game)**.** Set $N \in \mathbb{N}$. For each $t = 0, \dots, T-1$, let $\mathfrak{P}_t : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \twoheadrightarrow \mathfrak{P}_t(s_t, a_t, \mu_t) \subseteq \mathcal{P}(S)$ be the correspondence at time t given in Definition 2.1. Then an N agent Makov game under model uncertainty, say $(S, A, \mu^o, \mathfrak{P}^N_{0:T}, r \mid N, \mathfrak{P}_{0:T})$, comprises the following:

(i) $(S, \mathcal{B}_S)$ and $(A, \mathcal{B}_A)$ are Borel spaces for the state and action spaces, respectively.

(ii) $s^1_0, \dots, s^N_0$ are independent and identically distributed according to $\mu^o \in \mathcal{P}(S)$. Furthermore, denote by $\overline{\mu}^{o,N}(d\overline{s}^N_0) := \prod_{i=1}^N \mu^o(ds^i_0) \in \mathcal{P}(S^N)$.

(iii) For every $t = 0, \dots, T-1$, set $\mathfrak{P}_t^N : S^N \times A^N \ni (\overline{s}_t^N, \overline{a}_t^N) \twoheadrightarrow \mathfrak{P}_t^N(\overline{s}_t^N, \overline{a}_t^N) \subseteq \mathcal{P}(S^N)$ to be a correspondence at time t so that for every $(\overline{s}_t^N, \overline{a}_t^N) \in S^N \times A^N$,

$$\mathfrak{P}_t^N(\overline{s}_t^N, \overline{a}_t^N) := \left\{ \overline{\mathbb{P}}_t^N(d\overline{s}_{t+1}^N) := \prod_{i=1}^N \mathbb{P}_t^i(ds_{t+1}^i) \;\middle|\; \begin{array}{l} \text{for every } i = 1, \dots, N, \\ \mathbb{P}_t^i(ds_{t+1}^i) \in \mathfrak{P}_t(s_t^i, a_t^i, e^N(\overline{s}_t^N)) \end{array} \right\},$$

where $e^N(\cdot)$ is given in (2.5).

(iv) $r : S \times A \times S \times \mathcal{P}(S) \mapsto \mathbb{R}$ is a one-step Borel-measurable reward function.

Next, we introduce the set of Markov policies for the multi-agent model given in Definition 2.4 and the set of probability measures that induce model uncertainty in the underlying Markov game.

Definition 2.5. Given $N \in \mathbb{N}$, let $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$ be given in Definition 2.4.

(i) Denote by Π^N the N-tuple of sequences of Markov policies $\overline{\pi}_{0:T}^N := \prod_{i=1}^N \pi_{0:T}^i$ defined for every $t = 0, \dots, T-1$ by

$$\overline{\pi}_t^N : S^N \ni \overline{s}_t^N \mapsto \overline{\pi}_t^N(d\overline{a}_t^N | \overline{s}_t^N) := \prod_{i=1}^N \pi_t^i(da_t^i | s_t^i) \in \mathcal{P}(A^N),$$

where for every $i = 1, \dots, N$, $\pi_t^i : S \mapsto \mathcal{P}(A)$ denotes the Markov policy of agent i at time t.

(ii) Given $\overline{\pi}_{0:T}^N \in \Pi^N$, define by $\mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T}^N) \subseteq \mathcal{P}(S^N \times (S^N \times A^N)^T)$ the subset of all probability measures $\overline{\mathbb{P}}^N := \overline{\mu}^{o,N} \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}_0^N, \overline{p}_0^N)} \otimes \cdots \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}_{T-1}^N, \overline{p}_{T-1}^N)}$ such that for $t = 0, \dots, T-1$,

$$\overline{\mathbb{P}}^N_{(\overline{\pi}_t^N, \overline{p}_t^N)} : S^N \ni \overline{s}_t^N \mapsto \overline{\mathbb{P}}^N_{(\overline{\pi}_t^N, \overline{p}_t^N)}(d\overline{s}_{t+1}^N, d\overline{a}_t^N | \overline{s}_t^N) := \overline{p}_t^N(d\overline{s}_{t+1}^N | \overline{s}_t^N, \overline{a}_t^N) \overline{\pi}_t^N(d\overline{a}_t^N | \overline{s}_t^N)$$

is a stochastic kernel on $S^N \times A^N$ given S^N, where $\overline{p}_t^N : S^N \times A^N \mapsto \mathcal{P}(S^N)$ satisfies for every $(\overline{s}_t^N, \overline{a}_t^N) \in S^N \times A^N$ that

$$\overline{p}_t^N(d\overline{s}_{t+1}^N | \overline{s}_t^N, \overline{a}_t^N) := \prod_{i=1}^N p_t^i(ds_{t+1}^i | \overline{s}_t^N, \overline{a}_t^N) \in \mathfrak{P}_t^N(\overline{s}_t^N, \overline{a}_t^N)$$

with corresponding stochastic kernels $p_t^i : S^N \times A^N \mapsto \mathcal{P}(S)$, $i = 1, \dots, N$.

Having completed the description of the multi-agent Markov game under model uncertainty, we can proceed to describe the worst-case objective function of the individual agent: Given $N \in \mathbb{N}$, the worst-case objective function $J_i^N : \mathcal{P}(S) \times \Pi^N \ni (\mu^o, \overline{\pi}_{0:T}^N) \mapsto J_i^N(\mu^o, \overline{\pi}_{0:T}^N) \in \mathbb{R}$ of agent i, $i \in \{1, \dots, N\}$, is given by

$$J_i^N(\mu^o, \overline{\pi}_{0:T}^N) := \inf_{\overline{\mathbb{P}}^N \in \mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T}^N)} \mathbb{E}^{\overline{\mathbb{P}}^N}\left[\sum_{t=0}^{T-1} r\big(s_t^i, a_t^i, s_{t+1}^i, e^N(\overline{s}_t^N)\big)\right]. \tag{2.6}$$

Finally, we introduce the notion of a Markov-Nash equilibrium for the multi-agent Markov game under model uncertainty.

Definition 2.6 (Markov-Nash equilibria)**.** Given $N \in \mathbb{N}$, we say $(\pi_{0:T}^{*,1}, \dots, \pi_{0:T}^{*,N})$ is a Markov-Nash equilibrium of the N agent Markov game $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$ (see Definition 2.4) if $\overline{\pi}_{0:T}^{N|*} := \prod_{i=1}^N \pi_{0:T}^{*,i} \in \Pi^N$ satisfies that[3] for every $i = 1, \dots, N$

$$J_i^N(\mu^o, \overline{\pi}_{0:T}^{N|*}) = \sup_{\pi_{0:T} \in \Pi} J_i^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-i}, \pi_{0:T})).$$

[3] Denote by $(\overline{\pi}_{0:T}^{N|*,-i}, \pi_{0:T}) \in \Pi^N$ for every $t = 0, \dots, T-1$,

$$(\overline{\pi}_t^{N|*,-i}, \pi_t) := \pi_t(da_t^i | s_t^i) \prod_{j=1, j \neq i}^N \pi_t^{*,j}(da_t^j | s_t^j).$$

Moreover, for a given $\varepsilon > 0$, we say $(\pi^{*,1}_{0:T}, \ldots, \pi^{*,N}_{0:T})$ is an ε-Markov-Nash equilibrium of the N agent Markov game $(S, A, \mu^o, \mathfrak{P}^N_{0:T}, r \mid N, \mathfrak{P}_{0:T})$ if $\overline{\pi}^{N|*}_{0:T} \in \Pi^N$ satisfies for every $i = 1, \ldots, N$ that

$$J^N_i(\mu^o, \overline{\pi}^{N|*}_{0:T}) + \varepsilon \geq \sup_{\pi_{0:T} \in \Pi} J^N_i(\mu^o, (\overline{\pi}^{N|*,-i}_{0:T}, \pi_{0:T})).$$

3. Main results

3.1. Dynamic programming.

We first present some tailored dynamic programming results that will be useful for proving the existence of a mean-field equilibrium under model uncertainty.

Assumption 3.1. $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ given in Definition 2.1 satisfies the following conditions:

(i) S and A are finite subsets of a (possibly different) Euclidean space.

(ii) For every $t = 0, \ldots, T-1$, $\mathfrak{P}_t$ is non-empty, convex-valued, compact-valued, and continuous.[4] Furthermore, there exists a constant $C_{\mathfrak{P}_t} > 0$ such that for every $s_t \in S$, $a_t \in A$, $\mu_t, \tilde{\mu}_t \in \mathcal{P}(S)$ and for every $\mathbb{P} \in \mathfrak{P}_t(s_t, a_t, \mu_t)$, there exists $\tilde{\mathbb{P}} \in \mathfrak{P}_t(s_t, a_t, \tilde{\mu}_t)$ satisfying $d_{W_1}(\mathbb{P}, \tilde{\mathbb{P}}) \leq C_{\mathfrak{P}_t} d_{W_1}(\mu_t, \tilde{\mu}_t)$.

(iii) r is bounded and Lipschitz continuous in $\mathcal{P}(S)$, in the sense that there exists some constant $C_r > 0$, $L_r > 0$ such that for every $s_t, s_{t+1} \in S$, $a_t \in A$, and $\mu_t, \tilde{\mu}_t \in \mathcal{P}(S)$, $|r(s_t, a_t, s_{t+1}, \mu_t)| \leq C_r$ and $|r(s_t, a_t, s_{t+1}, \mu_t) - r(s_t, a_t, s_{t+1}, \tilde{\mu}_t)| \leq L_r d_{W_1}(\mu_t, \tilde{\mu}_t)$.

Let us formulate a sequence of auxiliary mappings $\widehat{V}_{0:T}$ backwards recursively as follows: for $t = T-1, \ldots, 0$, define $\widehat{V}_t : S \times (\mathcal{P}(S))^{T-t} \mapsto \mathbb{R}$ by setting for every $(s_t, \mu_{t:T}) \in S \times (\mathcal{P}(S))^{T-t}$

$$\widehat{V}_t(s_t, \mu_{t:T}) := \sup_{\pi \in \mathcal{P}(A)} \int_A \widehat{J}_t(s_t, a_t, \mu_{t:T}) \pi(da_t), \tag{3.1}$$

where $\widehat{J}_t : S \times A \times (\mathcal{P}(S))^{T-t} \mapsto \mathbb{R}$ is defined as follows: for every $(s_{T-1}, a_{T-1}, \mu_{T-1}) \in S \times A \times \mathcal{P}(S)$

$$\widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \mu_{T-1}) := \inf_{\mathbb{P} \in \mathfrak{P}_{T-1}(s_{T-1}, a_{T-1}, \mu_{T-1})} \int_S r(s_{T-1}, a_{T-1}, s_T, \mu_{T-1}) \mathbb{P}(ds_T), \tag{3.2}$$

whereas if $t \leq T-2$, we set for every $(s_t, a_t, \mu_{t:T}) \in S \times A \times (\mathcal{P}(S))^{T-t}$

$$\widehat{J}_t(s_t, a_t, \mu_{t:T}) := \inf_{\mathbb{P} \in \mathfrak{P}_t(s_t, a_t, \mu_t)} \int_S \Big(r(s_t, a_t, s_{t+1}, \mu_t) + \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}) \Big) \mathbb{P}(ds_{t+1}), \tag{3.3}$$

with $\mathfrak{P}_{0:T}$ given in Definition 2.1.

Finally, we define $\widehat{V} : (\mathcal{P}(S))^T \mapsto \mathbb{R}$ by setting for every $\mu_{0:T} \in (\mathcal{P}(S))^T$

$$\widehat{V}(\mu_{0:T}) := \int_S \widehat{V}_0(s_0, \mu_{0:T}) \mu_0(ds_0). \tag{3.4}$$

Lemma 3.2. *Suppose that Assumption 3.1 is satisfied. Let $\widehat{V}_{0:T}$ and $\widehat{J}_{0:T}$ be given in (3.1) and (3.2)–(3.3), respectively. Then the following statements hold for every $t = 0, \ldots, T-1$.*

(i) (Minimizer of $\widehat{J}_t$) There exists a measurable selector

$$\widehat{p}_t \colon S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \widehat{p}_t(\cdot | s_t, a_t, \mu_{t:T}) \in \mathfrak{P}_t(s_t, a_t, \mu_t)$$

satisfying that if $t = T-1$, then for every $(s_{T-1}, a_{T-1}, \mu_{T-1}) \in S \times A \times \mathcal{P}(S)$

$$\widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \mu_{T-1}) = \int_S r(s_{T-1}, a_{T-1}, s_T, \mu_{T-1}) \widehat{p}_{T-1}(ds_T | s_{T-1}, a_{T-1}, \mu_{T-1}), \tag{3.5}$$

[4] A correspondence between topological spaces is continuous if it is both lower- and upper-hemicontinuous (see, e.g., [2, Definition 17.2, p. 558]).

whereas if $t \le T-2$, then for every $(s_t, a_t, \mu_{t:T}) \in S \times A \times (\mathcal{P}(S))^{T-t}$

$$\widehat{J}_t(s_t, a_t, \mu_{t:T}) = \int_S \Big(r(s_t, a_t, s_{t+1}, \mu_t) + \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}) \Big) \widehat{p}_t(ds_{t+1}|s_t, a_t, \mu_{t:T}). \tag{3.6}$$

(ii) *(Maximizer of $\widehat{V}_t$) There exists a measurable selector*

$$\widehat{\pi}_t : S \times (\mathcal{P}(S))^{T-t} \ni (s_t, \mu_{t:T}) \mapsto \widehat{\pi}_t(\cdot|s_t, \mu_{t:T}) \in \mathcal{P}(A)$$

satisfying that for every $(s_t, \mu_{t:T}) \in S \times (\mathcal{P}(S))^{T-t}$

$$\widehat{V}_t(s_t, \mu_{t:T}) = \int_A \widehat{J}_t(s_t, a_t, \mu_{t:T}) \widehat{\pi}_t(da_t|s_t, \mu_{t:T}). \tag{3.7}$$

Remark 3.3. Berge's maximum theorem (see, e.g., [2, Theorem 17.31]), as presented in the proof of Lemma 3.2, ensures the existence of measurable selectors $\widehat{p}_{0:T}$ and $\widehat{\pi}_{0:T}$, as well as the following under the assumption therein: for every $t \le T-2$, the correspondence $\widehat{\mathfrak{P}}_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \twoheadrightarrow \widehat{\mathfrak{P}}_t(s_t, a_t, \mu_{t:T}) \subseteq \mathcal{P}(S)$ defined by

$$\widehat{\mathfrak{P}}_t(s_t, a_t, \mu_{t:T}) := \left\{ \mathbb{P} \in \mathfrak{P}_t(s_t, a_t, \mu_t) \;\middle|\; \begin{aligned} &\int_S \Big(r(s_t, a_t, s_{t+1}, \mu_t) + \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}) \Big) \mathbb{P}(ds_{t+1}) \\ &\quad = \widehat{J}_t(s_t, a_t, \mu_{t:T}) \end{aligned} \right\}$$

is non-empty, compact-valued, and upper-hemicontinuous (see [2, Theorem 17.31 (2.), (3.)]). Furthermore, since $\mathfrak{P}_t$ is convex-valued (see Assumption 3.1 (ii)), so is $\widehat{\mathfrak{P}}_t$. These observations will be used in Section 3.2.

As a consequence of Lemma 3.2, we obtain the following dynamic programming principle result.

Proposition 3.4. *Suppose that Assumption 3.1 is satisfied. Let $\widehat{V}_{0:T}$ and $\widehat{J}_{0:T}$ be given in (3.1) and (3.2)–(3.3), respectively. Given $\tilde{\mu}_{0:T} \in (\mathcal{P}(S))^T$, the following hold for every $t = 0, \dots, T-1$:*

(i) *There exists a stochastic kernel $p_t^* : S \times A \times \mathcal{P}(S) \mapsto \mathcal{P}(S)$ so that if $t = T-1$, then for every $(s_{T-1}, a_{T-1}) \in S \times A$*

$$\widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \tilde{\mu}_{T-1}) = \int_S r(s_{T-1}, a_{T-1}, s_T, \tilde{\mu}_{T-1}) p_{T-1}^*(ds_T|s_{T-1}, a_{T-1}, \tilde{\mu}_{T-1}), \tag{3.8}$$

whereas if $t \le T-2$, then for every $(s_t, a_t) \in S \times A$

$$\widehat{J}_t(s_t, a_t, \tilde{\mu}_{t:T}) = \int_S \Big(r(s_t, a_t, s_{t+1}, \tilde{\mu}_t) + \widehat{V}_{t+1}(s_{t+1}, \tilde{\mu}_{t+1:T}) \Big) p_t^*(ds_{t+1}|s_t, a_t, \tilde{\mu}_t). \tag{3.9}$$

Furthermore, there exists a Markov policy $\pi_t^* : S \mapsto \mathcal{P}(A)$ so that for every $s_t \in S$

$$\widehat{V}_t(s_t, \tilde{\mu}_{t:T}) = \int_A \widehat{J}_t(s_t, a_t, \tilde{\mu}_{t:T}) \pi_t^*(da_t|s_t). \tag{3.10}$$

(ii) *Let $p_{0:T}^*$ and $\pi_{0:T}^*$ be defined as in (i). Define $\mathbb{P}^*(\tilde{\mu}_{0:T}) \in \mathcal{Q}(\tilde{\mu}_{0:T}, \pi_{0:T}^*)$ by*

$$\mathbb{P}^*(\tilde{\mu}_{0:T}) := \mathbb{P}(\tilde{\mu}_{0:T}, \pi_{0:T}^*, p_{0:T}^*) := \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi_0^*, p_0^*)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi_{T-1}^*, p_{T-1}^*)}.$$

Then $V(\tilde{\mu}_{0:T})$ given in (2.2) is equal to $\widehat{V}(\tilde{\mu}_{0:T})$ given in (3.4), and $(\pi_{0:T}^*, p_{0:T}^*)$ are optimal for $V(\tilde{\mu}_{0:T})$, i.e.,

$$\begin{aligned} V(\tilde{\mu}_{0:T}) = J(\tilde{\mu}_{0:T}, \pi_{0:T}^*) &= \sup_{\pi_{0:T} \in \Pi} \mathbb{E}^{\mathbb{P}(\tilde{\mu}_{0:T}, \pi_{0:T}, p_{0:T}^*)} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \tilde{\mu}_t) \right] \\ &= \mathbb{E}^{\mathbb{P}^*(\tilde{\mu}_{0:T})} \left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \tilde{\mu}_t) \right], \end{aligned} \tag{3.11}$$

with $\mathbb{P}(\tilde{\mu}_{0:T}, \pi_{0:T}, p_{0:T}^*) := \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi_0, p_0^*)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi_{T-1}, p_{T-1}^*)} \in \mathcal{Q}(\tilde{\mu}_{0:T}, \pi_{0:T})$.

The proofs of Lemma 3.2 and Proposition 3.4 are presented in Section 5.1.

Next, we revisit the multi-agent Markov game given in Definitions 2.4 and 2.5 to obtain the corresponding dynamic programming principle result. This will be useful in Section 3.3 for determining the worst-case transition kernel for any given Markov policy.

Set $N \in \mathbb{N}$, and let $\overline{\pi}^N_{0:T} \in \Pi^N$ and $i \in \{1,\dots,N\}$. Define a sequence of mappings $\widehat{J}^N_{0:T,i}(\cdot,\cdot;\overline{\pi}^N_{0:T})$ backwards recursively as follows: Define for every $(\overline{s}^N_{T-1},\overline{a}^N_{T-1}) \in S^N \times A^N$

$$\widehat{J}^N_{T-1,i}(\overline{s}^N_{T-1},\overline{a}^N_{T-1};\overline{\pi}^N_{0:T}) := \inf_{\mathbb{P}\in\mathfrak{P}^N_{T-1}(\overline{s}^N_{T-1},\overline{a}^N_{T-1})} \int_{S^N} r\big(s^i_{T-1},a^i_{T-1},s^i_T,e^N(\overline{s}^N_{T-1})\big)\mathbb{P}(d\overline{s}^N_T), \tag{3.12}$$

and for $t \le T-2$, define for every $(\overline{s}^N_t,\overline{a}^N_t) \in S^N \times A^N$

$$\begin{aligned} \widehat{J}^N_{t,i}(\overline{s}^N_t,\overline{a}^N_t;\overline{\pi}^N_{0:T}) := \inf_{\mathbb{P}\in\mathfrak{P}^N_t(\overline{s}^N_t,\overline{a}^N_t)} \int_{S^N} \Bigg(& r\big(s^i_t,a^i_t,s^i_{t+1},e^N(\overline{s}^N_t)\big) \\ & + \int_{A^N} \widehat{J}^N_{t+1,i}(\overline{s}^N_{t+1},\overline{a}^N_{t+1};\overline{\pi}^N_{0:T})\overline{\pi}^N_{t+1}(d\overline{a}^N_{t+1}|\overline{s}^N_{t+1})\Bigg)\mathbb{P}(d\overline{s}^N_{t+1}), \end{aligned} \tag{3.13}$$

with $\mathfrak{P}^N_{0:T}$ given in Definition 2.4.

Lemma 3.5. *Suppose that Assumption 3.1 is satisfied. Set $N \in \mathbb{N}$, and let $\overline{\pi}^N_{0:T} \in \Pi^N$ and $i \in \{1,\dots,N\}$. Furthermore, let $\widehat{J}^N_{0:T,i}(\cdot,\cdot;\overline{\pi}^N_{0:T})$ be given in (3.12) and (3.13). Then for every $t = 0,\dots,T-1$, there exists a measurable selector (i.e., minimizer for $\widehat{J}^N_{t,i}(\cdot,\cdot;\overline{\pi}^N_{0:T})$)*

$$\widehat{p}_{(t,i,\overline{\pi}^N_{0:T})} : S^N \times A^N \ni (\overline{s}^N_t,\overline{a}^N_t) \mapsto \widehat{p}_{(t,i,\overline{\pi}^N_{0:T})}(\cdot|\overline{s}^N_t,\overline{a}^N_t) \in \mathfrak{P}^N_t(\overline{s}^N_t,\overline{a}^N_t)$$

satisfying that if $t = T-1$, then for every $(s^N_{T-1},a^N_{T-1}) \in S^N \times A^N$

$$\widehat{J}^N_{T-1,i}(\overline{s}^N_{T-1},\overline{a}^N_{T-1};\overline{\pi}^N_{0:T}) = \int_{S^N} r\big(s^i_{T-1},a^i_{T-1},s^i_T,e^N(\overline{s}^N_{T-1})\big)\widehat{p}_{(T-1,i,\overline{\pi}^N_{0:T})}(d\overline{s}^N_T|\overline{s}^N_{T-1},\overline{a}^N_{T-1}),$$

whereas if $t \le T-2$, then for every $(\overline{s}^N_t,\overline{a}^N_t) \in S^N \times A^N$

$$\begin{aligned} \widehat{J}^N_{t,i}(\overline{s}^N_t,\overline{a}^N_t;\overline{\pi}^N_{0:T}) = \int_{S^N} \Bigg(& r\big(s^i_t,a^i_t,s^i_{t+1},e^N(\overline{s}^N_t)\big) \\ & + \int_{A^N} \widehat{J}^N_{t+1,i}(\overline{s}^N_{t+1},\overline{a}^N_{t+1};\overline{\pi}^N_{0:T})\overline{\pi}^N_{t+1}(d\overline{a}^N_{t+1}|\overline{s}^N_{t+1})\Bigg)\widehat{p}_{(t,i,\overline{\pi}^N_{0:T})}(d\overline{s}^N_{t+1}|\overline{s}^N_t,\overline{a}^N_t). \end{aligned}$$

As a consequence of Lemma 3.5, we obtain the following result.

Proposition 3.6. *Suppose that Assumption 3.1 is satisfied. For every $i \in \{1,\dots,N\}$, initial distribution μ^o, and $\overline{\pi}^N_{0:T} \in \Pi^N$, let $\widehat{p}_{(0:T,i,\overline{\pi}^N_{0:T})}$ be given in Lemma 3.5. Then*

$$\overline{\mathbb{P}}^N(\mu^o,\overline{\pi}^N_{0:T},\widehat{p}_{(0:T,i,\overline{\pi}^N_{0:T})}) := \overline{\mu}^{o,N} \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}^N_0,\widehat{p}_{(0,i,\overline{\pi}^N_{0:T})})} \otimes \cdots \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}^N_{T-1},\widehat{p}_{(T-1,i,\overline{\pi}^N_{0:T})})} \in \mathcal{Q}^N(\mu^o,\overline{\pi}^N_{0:T})$$

is the worst-case measure for $J^N_i(\mu^o,\overline{\pi}^N_{0:T})$ (given in (2.6)), i.e.,

$$\begin{aligned} J^N_i(\mu^o,\overline{\pi}^N_{0:T}) &= \mathbb{E}^{\overline{\mathbb{P}}^N(\mu^o,\overline{\pi}^N_{0:T},\widehat{p}_{(0:T,i,\overline{\pi}^N_{0:T})})}\left[\sum_{t=0}^{T-1} r\big(s^i_t,a^i_t,s^i_{t+1},e^N(\overline{s}^N_t)\big)\right] \\ &= \int_{S^N}\int_{A^N} \widehat{J}^N_{0,i}(\overline{s}^N_0,\overline{a}^N_0)\overline{\pi}^N_0(d\overline{a}^N_0|\overline{s}^N_0)\overline{\mu}^{o,N}(d\overline{s}^N_0), \end{aligned}$$

with $\widehat{J}^N_{0,i}$ given in (3.13).

The proofs of Lemma 3.5 and Proposition 3.6 can be found in Section 5.2.

3.2. **Existence of mean-field equilibrium.** Using the results of the dynamic programming principle derived for the mean-field Markov game in Section 3.1, along with Kakutani's fixed point theorem (see, e.g., [2, Corollary 17.55, p. 583]), we will demonstrate the existence of a mean-field equilibrium under model uncertainty in Theorem 3.10.

Definition 3.7. Set $\Xi := (\mathcal{P}(S\times A))^T$. For $\nu_{0:T}\in\Xi$ and $t=0,\dots,T-1$, denote by $\nu_{t,S}$ the marginal of $\nu_t\in\mathcal{P}(S\times A)$ on S, i.e., $\nu_{t,S}(\cdot):=\nu_t(\cdot\times A)\in\mathcal{P}(S)$. Furthermore, denote by

$$\pi_t^{\nu}: S\ni s_t\mapsto \pi_t^{\nu}(\cdot|s_t)\in\mathcal{P}(A)$$

the disintegrating kernel of ν_t with respect to $\nu_{t,S}$, i.e., $\nu_t(ds_t,da_t)=\pi_t^{\nu}(da_t|s_t)\nu_{t,S}(ds_t)$.

Definition 3.8. Let Ξ be given in Definition 3.7. Let $\widehat{\mathfrak{P}}_{0:T-1}$ be given in Remark 3.3. Furthermore, let $\widehat{J}_{0:T}$ be given in (3.2) and (3.3). Define the following correspondences:

(i) $\mathcal{C}:\Xi\ni\nu_{0:T}\twoheadrightarrow\mathcal{C}(\nu_{0:T})\subseteq\Xi$ is defined by

$$\begin{aligned}\mathcal{C}(\nu_{0:T}) := \Big\{\tilde{\nu}_{0:T}\in\Xi\ \Big|\ &\tilde{\nu}_{0,S}=\mu^o \text{ and for every } t=0,\dots,T-2,\ \text{ there exists}\\ &p_t^{\tilde{\nu}}: S\times A\times(\mathcal{P}(S))^{T-t}\ni(s_t,a_t,\mu_{t:T})\mapsto p_t^{\tilde{\nu}}(\cdot|s_t,a_t,\mu_{t:T})\in\mathcal{P}(S)\\ &\text{s.t. for every } (s_t,a_t)\in S\times A,\ p_t^{\tilde{\nu}}(\cdot|s_t,a_t,\nu_{t:T,S})\in\widehat{\mathfrak{P}}_t(s_t,a_t,\nu_{t:T,S})\\ &\text{and } \tilde{\nu}_{t+1,S}(\cdot)=\int_{S\times A}p_t^{\tilde{\nu}}(\cdot|s_t,a_t,\nu_{t:T,S})\nu_t(ds_t,da_t)\Big\},\end{aligned}$$

and $\mathcal{B}:\Xi\ni\nu_{0:T}\twoheadrightarrow\mathcal{B}(\nu_{0:T})\subseteq\Xi$ is defined by

$$\mathcal{B}(\nu_{0:T}) := \Big\{\tilde{\nu}_{0:T}\in\Xi\ \Big|\ \text{for every } t=0,\dots,T-1,\ \tilde{\nu}_t(D_t(\nu_{t:T}))=1\Big\},$$

where $D_t(\nu_{t:T}):=\big\{(s_t,a_t)\in S\times A\mid \max_{a_t'\in A}\widehat{J}_t(s_t,a_t',\nu_{t:T,S})=\widehat{J}_t(s_t,a_t,\nu_{t:T,S})\big\}$.

(ii) $\Gamma:\Xi\ni\nu_{0:T}\twoheadrightarrow\Gamma(\nu_{0:T})\subseteq\Xi$ is defined by

$$\Gamma(\nu_{0:T}):=\mathcal{C}(\nu_{0:T})\cap\mathcal{B}(\nu_{0:T}).$$

We say $\nu_{0:T}\in\Xi$ is a *fixed point* of Γ if $\nu_{0:T}\in\Gamma(\nu_{0:T})$.

Proposition 3.9. *Suppose that Assumption 3.1 is satisfied. Then the following hold:*

(i) The correspondence Γ given in Definition 3.8 (ii) is non-empty and convex-valued.

(ii) The graph of Γ, i.e. $\mathrm{Gr}(\Gamma):=\{(\nu_{0:T},\xi_{0:T})\in\Xi\times\Xi\mid\xi_{0:T}\in\Gamma(\nu_{0:T})\}$, is closed.

(iii) There exists a fixed point $\nu_{0:T}^*\in\Xi$ of Γ, i.e., $\nu_{0:T}^*\in\Gamma(\nu_{0:T}^*)$.

Using a fixed point $\nu_{0:T}^*\in\Xi$ of Γ together with the measurable selectors given in Lemma 3.2, we obtain the following main theorem.

Theorem 3.10. *Let $(S,A,\mu^o,\mathfrak{P}_{0:T},r)$ be the mean-field Markov game under model uncertainty given in Definition 2.1. Suppose that Assumption 3.1 is satisfied. Then there exists a mean-field equilibrium $(\mu_{0:T}^*,\pi_{0:T}^*,p_{0:T}^*)$ of $(S,A,\mu^o,\mathfrak{P}_{0:T},r)$ (see Definition 2.3).*

The proofs of Proposition 3.9 and Theorem 3.10 can be found in Section 6.

3.3. **Existence of approximate Markov-Nash equilibrium.** Fix a mean-field equilibrium $(\mu_{0:T}^*,\pi_{0:T}^*,p_{0:T}^*)$ of the mean-field Markov game $(S,A,\mu^o,\mathfrak{P}_{0:T},r)$ (whose existence is ensured by Theorem 3.10 under the assumption therein).

In the following, we demonstrate that under certain assumptions, the optimal policy $\pi_{0:T}^*$ of the mean-field equilibrium constitutes an approximate Markov-Nash equilibrium of the multi-agent Markov game given in Definitions 2.4 and 2.5. To that end, we first introduce some key notions related to worst-case measures describing the multi-agent Markov game for a given policy.

Definition 3.11 (Worst-case measures)**.** Let $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. For every $N \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, we introduce the following.

(i) Denote by

$$\mathbb{P}^{*|(N)} := \mathbb{P}(\mu_{0:T}^*, \pi_{0:T}^{(N)}, p_{0:T}^*) \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^{(N)}),$$

where $\mathbb{P}(\mu_{0:T}^*, \pi_{0:T}^{(N)}, p_{0:T}^*)$ is given in (2.4). Moreover, if $\pi_{0:T}^{(N)} = \pi_{0:T}^*$, then we denote by

$$\mathbb{P}^* := \mathbb{P}^*(\mu_{0:T}^*) = \mathbb{P}^{*|(N)} \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^*)$$

the worst-case measure for $V(\mu_{0:T}^*)$ (see Proposition 3.4 (ii)).

(ii) For every $t \in \{0, \dots, T-1\}$, denote by

$$\overline{\pi}_{t,i}^{N|(N)} : S^N \ni \overline{s}_t^N \mapsto \overline{\pi}_{t,i}^{N|(N)}(d\overline{a}_t^N | \overline{s}_t^N) := \pi_t^{(N)}(da_t^i | s_t^i) \prod_{j=1, j\neq i}^{N} \pi_t^*(da_t^j | s_t^j),$$

$$\overline{p}_{t,i}^{N|(N)} : S^N \times A^N \ni (\overline{s}_t^N, \overline{a}_t^N) \mapsto \overline{p}_{t,i}^{N|(N)}(d\overline{s}_{t+1}^N | \overline{s}_t^N, \overline{a}_t^N) := \widehat{p}_{(t,i,\overline{\pi}_{0:T,i}^{N|(N)})}(d\overline{s}_{t+1}^N | \overline{s}_t^N, \overline{a}_t^N)$$

a Markov policy and a stochastic kernel, respectively, where $\widehat{p}_{(t,i,\overline{\pi}_{0:T,i}^{N|(N)})}$ is defined as in Lemma 3.5 with respect to $\overline{\pi}_{0:T,i}^{N|(N)}$. Moreover, let $\overline{\mathbb{P}}_i^{N|(N)} \in \mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T,i}^{N|(N)})$ be given by

$$\overline{\mathbb{P}}_i^{N|(N)} := \overline{\mathbb{P}}^N(\mu^o, \overline{\pi}_{0:T,i}^{N|(N)}, \overline{p}_{0:T,i}^{N|(N)}) := \overline{\mu}^{o,N} \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}_{0,i}^{N|(N)}, \overline{p}_{0,i}^{N|(N)})} \otimes \cdots \otimes \overline{\mathbb{P}}^N_{(\overline{\pi}_{T-1,i}^{N|(N)}, \overline{p}_{T-1,i}^{N|(N)})}$$

so that it is the worst-case measure for $J_i^N(\mu^o, \overline{\pi}_{0:T,i}^{N|(N)})$ given in (2.6) (see Proposition 3.6).

The notions introduced in the following, which elaborate on certain laws and stochastic kernels for the one-step reward function $r : S \times A \times S \times \mathcal{P}(S) \mapsto \mathbb{R}$ under the worst-case measures (described above), will be used in Propositions 3.16 and 3.17.

Definition 3.12 (Laws and kernels under worst-case measures)**.** Let $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. For every $N \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, we define the following: Let $\mathbb{P}^{*|(N)} \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^{(N)})$, $\mathbb{P}^* \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^*)$, and $\overline{\mathbb{P}}_i^{N|(N)} \in \mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T,i}^{N|(N)})$ be given in Definition 3.11. Then for every $t = 0, \dots, T-1$,

(i) Denote by

$$\mathbb{M}_t^{*|(N)}(ds_t, da_t) \in \mathcal{P}(S \times A), \qquad \mathbb{M}_{t,i}^{N|(N)}(ds_t, da_t) \in \mathcal{P}(S \times A)$$

the law of (s_t, a_t) under $\mathbb{P}^{*|(N)}$ and the law of (s_t^i, a_t^i) under $\overline{\mathbb{P}}_i^{N|(N)}$, respectively, at time t. Moreover, if $\pi_{0:T}^{(N)} = \pi_{0:T}^*$, then for every $t = 0, \dots, T-1$ set

$$\mathbb{M}_t^*(ds_t, da_t) := \mathbb{M}_t^{*|(N)}(ds_t, da_t) \in \mathcal{P}(S \times A)$$

to be the law of (s_t, a_t) under $\mathbb{P}^*$.

(ii) Denote by

$$\mathbb{K}_{t,i}^{N|(N)} : S \times A \ni (s_t, a_t) \mapsto \mathbb{K}_{t,i}^{N|(N)}(ds_{t+1}, d\mu_t | s_t, a_t) \in \mathcal{P}(S \times \mathcal{P}(S))$$

the stochastic kernel on $S \times \mathcal{P}(S)$ given $S \times A$ so that $\mathbb{K}_{t,i}^{N|(N)}(ds_{t+1}, d\mu_t | s_t, a_t)$ is the conditional law of $(s_{t+1}^i, e^N(\overline{s}_t^N))$ given $(s_t^i, a_t^i) = (s_t, a_t) \in S \times A$ under $\overline{\mathbb{P}}_i^{N|(N)}$ at time t.

(iii) Let $\mathbb{Q}_t^{*|(N)}, \mathbb{Q}_t^{N|(N)} \in \mathcal{P}(S \times A \times S \times \mathcal{P}(S))$ be given by[5]

$$\mathbb{Q}_t^{*|(N)}(ds_t, da_t, ds_{t+1}, d\mu_t) := p_t^*(ds_{t+1} | s_t, a_t, \mu_t)\; \delta_{\mu_t^*}(d\mu_t)\; \mathbb{M}_t^{*|(N)}(ds_t, da_t),$$

$$\mathbb{Q}_{t,i}^{N|(N)}(ds_t, da_t, ds_{t+1}, d\mu_t) := \mathbb{K}_{t,i}^{N|(N)}(ds_{t+1}, d\mu_t | s_t, a_t)\; \mathbb{M}_{t,i}^{N|(N)}(ds_t, da_t),$$

[5] Denote by $\delta_{\mu_t^*} \in \mathcal{P}(\mathcal{P}(S))$ the Dirac measure on $\mathcal{P}(S)$ at $\mu_t^* \in \mathcal{P}(S)$.

so that

- $\mathbb{Q}_t^{*|(N)}$ is the law of $(s_t, a_t, s_{t+1}, \mu_t)$ under $\mathbb{P}^{*|(N)}$ at time t with $\mu_t = \mu_t^*$.
- $\mathbb{Q}_{t,i}^{N|(N)}$ is the law of $(s_t^i, a_t^i, s_{t+1}^i, e^N(\overline{s}_t^N))$ under $\overline{\mathbb{P}}_i^{N|(N)}$ at time t.

Moreover, if $\pi_{0:T}^{(N)} = \pi_{0:T}^*$, we let $\mathbb{Q}_t^* \in \mathcal{P}(S \times A \times S \times \mathcal{P}(S))$ be given by

$$\mathbb{Q}_t^* := \mathbb{Q}_t^{*|(N)}$$

so that it is the law of $(s_t, a_t, s_{t+1}, \mu_t)$ under $\mathbb{P}^*$ at time t with $\mu_t = \mu_t^*$.

In Remark 7.2 (see Section 7.1), we provide explicit characterizations for the laws and stochastic kernels described in Definition 3.12.

Remark 3.13. Let $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. For every $N \in \mathbb{N}$, by the definition of $\mathfrak{P}_{0:T}^N$ and $\overline{\pi}_{0:T,i}^{N|(N)}$ (given in Definition 2.4 (iii) and Definition 3.11 (ii), respectively), all of the laws $\mathbb{M}_{0:T,i}^{N|(N)}$ and kernels $\mathbb{K}_{0:T,i}^{N|(N)}$ (given in Definition 3.12 (i), (ii)) are identical for each $i \in \{1, \dots, N\}$. Consequently, all the laws $\mathbb{Q}_{0:T,i}^{N|(N)}$ are also identical. Therefore, for every $t = 0, \dots, T-1$ we simplify their notations as follows: for every $i = 1, \dots, N$

$$\mathbb{M}_t^{N|(N)} := \mathbb{M}_{t,i}^{N|(N)}, \qquad \mathbb{K}_t^{N|(N)} := \mathbb{K}_{t,i}^{N|(N)}, \qquad \mathbb{Q}_t^{N|(N)} := \mathbb{Q}_{t,i}^{N|(N)}.$$

We impose the following conditions on the stochastic kernels $\mathbb{K}_{0:T}^{N,N}$ given in Remark 3.13.

Assumption 3.14. For any $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$, the following holds: for every $t = 0, \dots, T-1$ and $(s_t, a_t) \in S \times A$, as $N \to \infty$,

$$\mathbb{K}_t^{N|(N)}(ds_{t+1}, d\mu_t | s_t, a_t) \rightharpoonup p_t^*(ds_{t+1}|s_t, a_t, \mu_t)\, \delta_{\mu_t^*}(d\mu_t),$$

where $(\mu_{0:T}^*, \pi_{0:T}^*, p_{0:T}^*)$ is the (fixed) mean-field equilibrium.

Remark 3.15. Under the Nash Certainty Equivalence Principle, the decentralized game without model uncertainty can be reduced to a single-agent decision (see, e.g., [34]). The state evolution of a representative agent should be consistent with the total population behavior. To extend this idea to our framework under model uncertainty, we need to ensure the following.

From an agent's perspective in $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$, under 'any' state and action, her behavior should converge to the representative agent's behavior in $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$. Additionally, the behavior of the rest of the population, modeled via the empirical distribution, should converge to the population's behavior in $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ (i.e., the state-measure flow $\mu_{0:T}^*$). For a sequence of arbitrary policies $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$, we observe that as $N \to \infty$, the influence of an individual agent's state and action on the overall population becomes increasingly negligible. Since every other agent follows the mean field equilibrium policy $\pi_{0:T}^*$ (see Definition 3.11 (ii)), the overall state distribution in $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$ should still converge to the state distribution in the mean-field equilibrium, regardless of the state and action the one individual agent might be in.

If the agent also chooses the mean-field equilibrium policy, i.e., $\pi_{0:T}^{(N)} := \pi_{0:T}^*$, we need to ensure that the state evolution of a representative agent is consistent with the total population behavior as $N \to \infty$. By the definition of the mean-field equilibrium given in Definition 2.3 (ii), we obtain such consistency exactly there. Hence, Assumption 3.14 guarantees that as N grows larger, both the individual and total population behaviors in $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$ converge to a state under which the Nash Certainty Equivalence Principle will hold.

Proposition 3.16 allows us to connect the expected one-step rewards of $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$ and $(S, A, \mu^o, \mathfrak{P}_{0:T}^N, r \mid N, \mathfrak{P}_{0:T})$ by using the laws and kernels given in Definition 3.12 and Remark 3.13.

Proposition 3.16. *Suppose that Assumptions 3.1 and 3.14 are satisfied. Let $(\pi^{(N)}_{0:T})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. Moreover, for every $N \in \mathbb{N}$, let $\mathbb{Q}^{*|(N)}_{0:T}, \mathbb{Q}^{N|(N)}_{0:T} \subseteq \mathcal{P}(S \times A \times S \times \mathcal{P}(S))$ be given in Definition 3.12 (iii) and Remark 3.13, respectively. Then for every $t = 0, \dots, T-1$, the following holds: for every $g \in C_b(S \times A \times S \times \mathcal{P}(S))$*

$$\lim_{N\to\infty} \left| \mathbb{E}^{\mathbb{Q}^{N|(N)}_t}\left[g(s_t, a_t, s_{t+1}, \mu_t)\right] - \mathbb{E}^{\mathbb{Q}^{*|(N)}_t}\left[g(s_t, a_t, s_{t+1}, \mu_t)\right] \right| = 0. \tag{3.14}$$

As a consequence, we obtain the following.

Proposition 3.17. *Suppose that Assumptions 3.1 and 3.14 are satisfied. Let $(\pi^{(N)}_{0:T})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. For every $N \in \mathbb{N}$, let $J^N_1(\mu^o, \overline{\pi}^{N|(N)}_{0:T,1})$ be the worst-case objective function of the agent 1 under $(\mu^o, \overline{\pi}^{N|(N)}_{0:T,1})$ (see Definition 3.11 (ii)) and let $\mathbb{P}^{*|(N)} \in \mathcal{Q}(\mu^*_{0:T}, \pi^{(N)}_{0:T})$ be given in Definition 3.11 (i). Then it holds that*

$$\lim_{N\to\infty} \left| J^N_1(\mu^o, \overline{\pi}^{N|(N)}_{0:T,1}) - \mathbb{E}^{\mathbb{P}^{*|(N)}}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu^*_t)\right] \right| = 0.$$

The proofs of Proposition 3.16 and 3.17 are presented in Section 7.1.

Remark 3.18. Since $V(\mu^*_{0:T}) = \mathbb{E}^{\mathbb{P}^*(\mu^*_{0:T})}[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu^*_t)]$ (see Proposition 3.4 (ii)),

$$\lim_{N\to\infty} J^N_1(\mu^o, \overline{\pi}^{N|*}_{0:T}) = V(\mu^*_{0:T})$$

follows directly from Proposition 3.17 (with $\overline{\pi}^{N|*}_{0:T}$ defined in (3.15)).

Combining Propositions 3.16 and 3.17 with the optimality of $\pi^*_{0:T}$ in the mean-field equilibrium (see Definition 2.3 (i)), we conclude in Theorem 3.19 that the Markov policy $\pi^*_{0:T}$ forms an approximate Markov-Nash equilibrium. The corresponding proof can be found in Section 7.2.

Theorem 3.19. *Suppose that Assumptions 3.1 and 3.14 are satisfied. Then for any given $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that for each $N \geq N(\varepsilon)$, $(\pi^*_{0:T}, \cdots, \pi^*_{0:T})$ is an ε-Markov-Nash equilibrium of $(S, A, \mu^o, \mathfrak{P}^N_{0:T}, r \mid N, \mathfrak{P}_{0:T})$ (see Definition 2.6), i.e., $\overline{\pi}^{N|*}_{0:T} \in \Pi^N$ defined for every $t = 0, \dots, T-1$ by*

$$\overline{\pi}^{N|*}_t : S^N \ni \overline{s}^N_t \mapsto \overline{\pi}^{N|*}_t(d\overline{a}^N_t | \overline{s}^N_t) := \textstyle\prod_{j=1}^N \pi^*_t(da^j_t | s^j_t) \tag{3.15}$$

satisfies that for every $i = 1, \dots, N$, $J^N_i(\mu^o, \overline{\pi}^{N|*}_{0:T}) + \varepsilon \geq \sup_{\pi_{0:T}\in\Pi} J^N_i(\mu^o, (\overline{\pi}^{N|*,-i}_{0:T}, \pi_{0:T}))$.

4. Numerical example: Crowd motion under model uncertainty

Based on Proposition 3.4 and Theorem 3.10, we derive an iterative scheme that allows to compute approximately a mean-field equilibrium $(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T})$ of $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$. We provide a pseudo-code in Algorithm 1 to show how it can be implemented.[6]

The algorithm proceeds as follows: Starting with given $\mu^*_{0:T} \in (\mathcal{P}(S))^T$, we apply the dynamic programming results as described in (3.1)–(3.4) to derive the worst-case kernels $p^*_{0:T}$ and optimal Markov policies $\pi^*_{0:T}$ for $V(\mu^*_{0:T})$ (see Proposition 3.4). Next, we update $\mu^*_{0:T}$ by constructing a new sequence of state measures in the sense of Definition 2.3 (ii). This process is iterated until we attain a fixed point $(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T})$ in the sense of Proposition 3.9 and Theorem 3.10. Note that as S and A are finite, in line with Assumption 3.1 (i), we will construct the corresponding probability measures by interpreting them as elements of a simplex in an Euclidean space.

[6] All the numerical experiments have been performed with the following hardware configurations: a Macbook Air with Apple M1 chip, 8 GBytes of memory, and Mac OS 13.0. All the codes are provided in the following link: https://github.com/JoLa2606/robust_MFE/

Algorithm 1 An iteretative scheme for mean-field equilibrium (MFE) under model uncertainty

```
Input: $(S, A)$ with $n(S), n(A) < \infty$ (satisfying Assumption 3.1 (i)), $\mu^o \in \mathcal{P}(S)$ (i.e., initial distribution),
        $(\mathfrak{P}_{0:T}, r)$ (satisfying Assumption 3.1 (ii), (iii)), and $\mu^*_{0:T} \in (\mathcal{P}(S))^T$ (a priori arbitrarily chosen);
Function MFE($\mu^*_{0:T}$; $S, A, \mu^o, \mathfrak{P}_{0:T}, r$):
    Set $\mu^*_0 := \mu^o$;
    while $\mu^*_{1:T}$ still changes
        for $t = T-1$ to $0$
            for $i = 1$ to $n(S)$
                for $j = 1$ to $n(A)$
                    if $t = T-1$
                        Compute $p^*_{T-1}(\cdot \mid s_i, a_j, \mu^*_{T-1}) \in \mathfrak{P}_{T-1}(s_i, a_j, \mu^*_{T-1})$ so that
                        $\widehat{J}_{T-1}(s_i, a_j, \mu^*_{T-1}) = \sum_{s \in S} r(s_i, a_j, s, \mu^*_{T-1}) p^*_{T-1}(s \mid s_i, a_j, \mu^*_{T-1})$;
                    else
                        Compute $p^*_t(\cdot \mid s_i, a_j, \mu^*_t) \in \mathfrak{P}_t(s_i, a_j, \mu^*_t)$ so that
                        $\widehat{J}_t(s_i, a_j, \mu^*_{t:T}) = \sum_{s \in S} \big(r(s_i, a_j, s, \mu^*_t) + \widehat{V}_{t+1}(s_{t+1}, \mu^*_{t+1:T})\big) p^*_t(s \mid s_i, a_j, \mu^*_t)$;
                end
                Compute $\pi^*_t(\cdot \mid s_i) \in \mathcal{P}(A)$ so that $\widehat{V}_t(s_i, \mu^*_{t:T}) = \sum_{a \in A} \widehat{J}_t(s_i, a, \mu^*_{t:T}) \pi^*_t(a|s_i)$;
            end
        end
        for $t = T-2$ to $0$
            Update $\mu^*_{t+1}$ so that $\mu^*_{t+1}(s_i) := \sum_{s \in S} \sum_{a \in A} p^*_t(s_i|s, a, \mu^*_t) \pi^*_t(a|s) \mu^*_t(s) \quad \forall i = 1, \dots, n(S)$;
        end
    end
    Return $(\mu^*_{0:T}, \pi^*_{0:T}, p^*_{0:T})$
```

We consider the following model, which can be found in [41, Section 5.7] and is inspired by the model studied in [20], and extend it by allowing for model uncertainty.

Definition 4.1. Let $S := \{0, 1, \dots, 4\}$ and $A := \{-1, 0, 1\}$ be state and action spaces, respectively. Furthermore, let $T := 2$ be the time horizon, and let $\lambda \geq 0$ and $c > 0$ be given. Agents can decide to move along the one-dimensional (1D) grid world S in both directions or stay where they are; we model these actions by *left* $= -1$, *stay* $= 0$, or *right* $= 1$.

(i) For every $t = 0, 1$, define $\mathfrak{P}^\lambda_t : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \twoheadrightarrow \mathfrak{P}^\lambda_t(s_t, a_t, \mu_t) \subseteq \mathcal{P}(S)$ by

$$\mathfrak{P}^\lambda_t(s_t, a_t, \mu_t) := \Big\{ \mathbb{P} \in \mathcal{P}(S) \;\Big|\; d_{W_1}\big(\mathbb{P}, p^o(\cdot|s_t, a_t, \mu_t)\big) \leq \lambda \Big\},$$

where $d_{W_1}(\cdot, \cdot)$ is the 1-Wasserstein distance on S and $p^o \colon S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \mapsto p^o(\cdot|s_t, a_t, \mu_t) \in \mathcal{P}(S)$ is a reference stochastic kernel on S given $S \times A \times \mathcal{P}(S)$ so that under $p^o(\cdot|s_t, a_t, \mu_t)$, s_{t+1} satisfies

$$s_{t+1} = \begin{cases} s_t + a_t + \varepsilon_{t+1} & \text{if } \; s_t + a_t + \varepsilon_{t+1} \in S, \\ s_t & \text{else}, \end{cases}$$

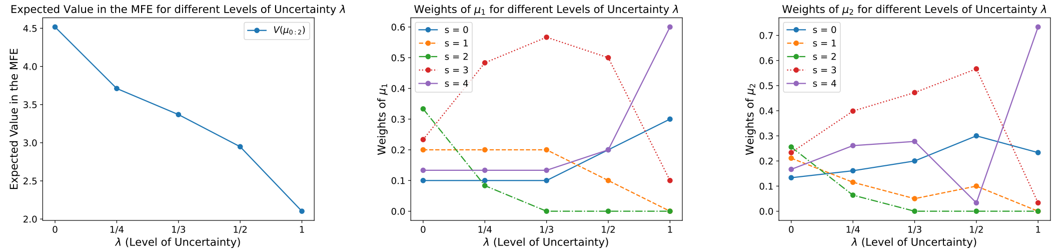
where ε_{t+1} is independently identically distributed according to a uniform distribution with values in A.

(ii) Define $r : S \times A \times S \times \mathcal{P}(S) \mapsto \mathbb{R}$ by setting for every $(s, a, \hat{s}, \mu) \in S \times A \times S \times \mathcal{P}(S)$,

$$r(s, a, \hat{s}, \mu) := \Big(1 - \frac{1}{2}|\hat{s} - 2|\Big) - \frac{|a|}{4} - \log\big(\mu(\hat{s}) + c\big).$$

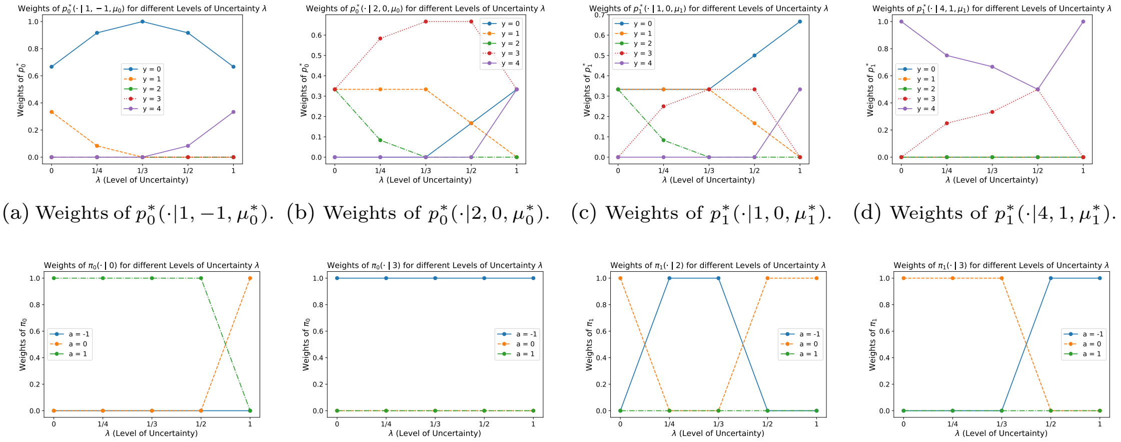
Lemma 4.2. *Under the setup given in Definition 4.1, let $\lambda \geq 0$ and $c > 0$ be given. Then, the set-valued maps $\mathfrak{P}^\lambda_{0:T}$ and the one-step reward function r satisfy Assumption 3.1 (ii) and (iii).*

The proof of the above lemma can be found in Appendix A.



(a) Values for $V(\mu^*_{0:2})$. (b) Weights of μ^*_1. (c) Weights of μ^*_2.

FIGURE 1. Sensitivity of values for $V(\mu^*_{0:2})$ and weights of $\mu^*_{0:2}$ with respect to uncertainty level λ with given $\mu^o = \mu^*_0 = (0.2, 0.1, 0.05, 0.25, 0.4)$.



(a) Weights of $p^*_0(\cdot|1,-1,\mu^*_0)$. (b) Weights of $p^*_0(\cdot|2,0,\mu^*_0)$. (c) Weights of $p^*_1(\cdot|1,0,\mu^*_1)$. (d) Weights of $p^*_1(\cdot|4,1,\mu^*_1)$.

(e) Weights of $\pi^*_0(\cdot \mid 0)$. (f) Weights of $\pi^*_0(\cdot \mid 3)$. (g) Weights of $\pi^*_1(\cdot \mid 2)$. (h) Weights of $\pi^*_1(\cdot \mid 3)$.

FIGURE 2. Sensitivity of $(\pi^*_{0:2}, p^*_{0:2})$ with respect to uncertainty level λ.

Remark 4.3. The one-step reward r is designed to encourage the agent to move toward the center while avoiding overly crowded areas. Additionally, it discourages unnecessary movement unless it is beneficial. The parameter c allows to model the degree of aversion of crowds. According to the reference kernel p^o, the agent can either remain in her current position or move to one of the adjacent positions. Moreover, the random disturbance ε_{t+1} may influence the dynamics, representing scenarios such as a concert where people prefer to be near the center but also wish to avoid excessively crowded spots. Agents try to move around in front of the stage based on their own actions but can also be randomly pushed around by the crowd.

Explicitly, we fix $c = 10^{-7}$ and consider different levels of uncertainty $\lambda \in \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1\}$. Let $\mu^o = (w_0^{\mu^o}, \ldots, w_4^{\mu^o}) = (0.2, 0.1, 0.05, 0.25, 0.4)$ be the initial state distribution.

Fig. 1(a) shows that the expected value $V(\mu^*_{0:2})$ decreases as the uncertainty λ increases, which is expected since a higher uncertainty level entails a potentially worse scenario.

Examining the state-flow measure μ^*_1 at $t = 1$, in Fig. 1(b) we observe that in the absence of model uncertainty, the majority of the weight is concentrated at the center position $s = 2$, with some weight distributed to the adjacent positions $s = 1$ and $s = 3$. The least weight is found at the extreme positions $s = 0$ and $s = 4$. This distribution can be interpreted that most individuals move towards the center, while a few choose to remain at the sides to avoid overcrowding. Whereas if the level of uncertainty increases, the distribution shifts, resulting in more weight being moved away from the center and an increase in the mass at $s = 3$. With large uncertainty, the mass is almost entirely shifted to the boundaries, $s = 0$ and $s = 4$. Similar effects are observed in the state-flow measure μ_2^* at $t = 2$, as shown in Fig. 1(c).

Fig. 2 shows the sensitivity of the optimal pair $(\pi_{0:2}^*, p_{0:2}^*)$ in the mean-field equilibrium with respect to uncertainty level λ.

Although it is hard to interpret the sensitivity of the worst-case stochastic kernels $p_{0:2}^*$ shown in Fig. 2 (a)–(d), we can at least observe that our model uncertainty framework described in Definition 4.1 (i) is working non-trivially.

Without model uncertainty, i.e. $\lambda = 0$, the strategy π_0^* at time $t = 0$ makes the agent move to the center $s = 2$ as the center is not crowded yet, as shown in Fig. 2 (e), (f). Indeed, we have seen in Fig. 1(b) that the weight of μ_1^* at $s = 2$ is dominant. On the other hand, to avoid the crowd at time $t = 1$, it becomes beneficial to stay at $s = 3$ rather than trying to move to the center $s = 2$ while those already at the center remain there, as shown in Fig. 2 (g), (h).

As the uncertainty level increases, we observe some interesting effects. In Fig. 2 (a)–(d), similar developments are observed across all presented scenarios for the worst-case kernels p_0^* and p_1^*. With increasing uncertainty, the probability of getting shifted to overly crowded areas, particularly to $s = 4$, increases. In Fig. 2(e), the optimal strategy shifts from attempting to move towards the center, $s = 2$, to staying at $s = 0$, i.e., avoiding movement to the right. Fig. 2(g) shows a similar effect: although being in the center is highly beneficial, the optimal strategy $\pi_1^*(\cdot \mid 2)$ becomes to resist moving to the crowded areas ($s = 3$ and $s = 4$). In Fig. 2(h), to avoid staying in the overly crowded area $s = 3$ or moving to $s = 4$, $\pi_1^*(\cdot \mid 3)$ changes in order to try to move towards the center.

5. Proof of results in Section 3.1

5.1. Proof of Lemma 3.2 and Proposition 3.4.

Lemma 5.1. *Suppose that Assumption 3.1 is satisfied. Let $\widehat{V}_{0:T}$ be given in (3.1). Fix any $t \in \{0, 1, \ldots, T-2\}$ and assume that there exist some constants $\widehat{C}_{t+1} \geq 1$ and $\widehat{L}_{t+1} > 0$ such that for every $s_{t+1} \in S$ and every $\mu_{t+1:T}, \tilde{\mu}_{t+1:T} \in (\mathcal{P}(S))^{T-t-1}$, it holds that*

$$
\begin{aligned}
&|\widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T})| \leq \widehat{C}_{t+1}, \\
&\left|\widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}) - \widehat{V}_{t+1}(s_{t+1}, \tilde{\mu}_{t+1:T})\right| \leq \widehat{L}_{t+1} \sum_{u=t+1}^{T-1} d_{W_1}(\mu_u, \tilde{\mu}_u).
\end{aligned}
\tag{5.1}
$$

Then the following hold:

(i) $\widehat{J}_t$ given in (3.3) is continuous on $S \times A \times (\mathcal{P}(S))^{T-t}$. Furthermore, there exists a measurable selector $\widehat{p}_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \widehat{p}_t(\cdot|s_t, a_t, \mu_{t:T}) \in \mathfrak{P}_t(s_, a_t, \mu_t)$ satisfying (3.6).

(ii) There exists a constant $\widehat{K}_t > 0$ such that for every $s_t \in S$, $a_t \in A$, and every $\mu_{t:T}, \tilde{\mu}_{t:T} \in (\mathcal{P}(S))^{T-t}$, $|\widehat{J}_t(s_t, a_t, \mu_{t:T}) - \widehat{J}_t(s_t, a_t, \tilde{\mu}_{t:T})| \leq \widehat{K}_t \sum_{u=t}^{T-1} d_{W_1}(\mu_u, \tilde{\mu}_u)$.

(iii) $\widehat{V}_t$ is continuous on $S \times (\mathcal{P}(S))^{T-t}$. Furthermore, there exists a measurable selector $\widehat{\pi}_t : S \times \mathcal{P}(S)^{T-t} \ni (s_t, \mu_{t:T}) \mapsto \widehat{\pi}_t(\cdot|s_t, \mu_{t:T}) \in \mathcal{P}(A)$ satisfying (3.7).

(iv) *There exist some constants $\widehat{C}_t \geq 1$ and $\widehat{L}_t > 0$ such that for every $s_t \in S$ and every $\mu_{t:T}, \tilde{\mu}_{t:T} \in (\mathcal{P}(S))^{T-t}$,*

$$|\widehat{V}_t(s_t, \mu_{t:T})| \leq \widehat{C}_t, \qquad \left|\widehat{V}_t(s_t, \mu_{t:T}) - \widehat{V}_t(s_t, \tilde{\mu}_{t:T})\right| \leq \widehat{L}_t \sum_{u=t}^{T-1} d_{W_1}(\mu_u, \tilde{\mu}_u).$$

Proof. We start by proving (i). To that end, set

$$\mathcal{S} := \left\{(s_t, a_t, \mu_{t:T}, p_t) \middle| (s_t, a_t, \mu_{t:T}) \in S \times A \times (\mathcal{P}(S))^{T-t}, p_t \in \mathfrak{P}_t(s_t, a_t, \mu_t)\right\}$$

and define an auxiliary map $F : \mathcal{S} \ni (s_t, a_t, \mu_{t:T}, p_t) \mapsto F(s_t, a_t, \mu_{t:T}, p_t) \in \mathbb{R}$ by

$$F(s_t, a_t, \mu_{t:T}, p_t) := \int_S \Big(r(s_t, a_t, s_{t+1}, \mu_t) + \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T})\Big)\, p_t(ds_{t+1}).$$

Then we consider a sequence $(s_t^n, a_t^n, \mu_{t:T}^n, p_t^n)_{n\in\mathbb{N}} \subseteq \mathcal{S}$ such that $(s_t^n, a_t^n) \to (s_t^\star, a_t^\star)$, $\mu_u^n \rightharpoonup \mu_u^\star$ (for every $u = t, \ldots, T-1$), and $p_t^n \rightharpoonup p_t^\star$ as $n \to \infty$, with some $(s_t^\star, a_t^\star, \mu_{t:T}^\star, p_t^\star) \in \mathcal{S}$.

By the triangle inequality, for every $n \in \mathbb{N}$,

$$\begin{aligned}
&|F(s_t^n, a_t^n, \mu_{t:T}^n, p_t^n) - F(s_t^\star, a_t^\star, \mu_{t:T}^\star, p_t^\star)| \\
&\quad \leq |F(s_t^\star, a_t^\star, \mu_{t:T}^\star, p_t^n) - F(s_t^\star, a_t^\star, \mu_{t:T}^\star, p_t^\star)| + |F(s_t^n, a_t^n, \mu_{t:T}^n, p_t^n) - F(s_t^\star, a_t^\star, \mu_{t:T}^\star, p_t^n)| =: \mathrm{I}^n + \mathrm{II}^n\,.
\end{aligned}$$

We will show that I^n and II^n vanish as $n \to \infty$.

From Assumption 3.1 (i), (iii), and (5.1), it follows that $r(s_t^\star, a_t^\star, \cdot, \mu_t^\star) + \widehat{V}_{t+1}(\cdot, \mu_{t+1:T}^\star)$ are continuous and bounded in S, i.e., for every $s_{t+1} \in S$, $|r(s_t^\star, a_t^\star, s_{t+1}, \mu_t^\star) + \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}^\star)| \leq (C_r + \widehat{C}_{t+1})$. Furthermore, since $p_t^n \rightharpoonup p_t^\star$ as $n \to \infty$, we obtain that $\lim_{n\to\infty} \mathrm{I}^n = 0$.

It remains to show the limit of II^n. By Assumption 3.1 (i), S and A are finite. Hence, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $(s_t^n, a_t^n) = (s_t^\star, a_t^\star)$. By Assumption 3.1 (iii) and (5.1), for every $n \geq N$,

$$\begin{aligned}
\mathrm{II}^n &\leq \int_X \Bigg(\left|r(s_t^\star, a_t^\star, s_{t+1}, \mu_t^n) - r(s_t^\star, a_t^\star, s_{t+1}, \mu_t^\star)\right| \\
&\qquad\qquad + \left|\widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}^n) - \widehat{V}_{t+1}(s_{t+1}, \mu_{t+1:T}^\star)\right|\Bigg) p^n(ds_{t+1}) \\
&\leq L_r d_{W_1}(\mu_t^n, \mu_t^\star) + \widehat{L}_{t+1} \sum_{u=t+1}^{T-1} d_{W_1}(\mu_u^n, \mu_u^\star).
\end{aligned}$$

The limit $\mu_u^n \rightharpoonup \mu_u^\star$ (for every $u = t, \ldots, T-1$) ensures that II^n vanishes as $n \to \infty$. Therefore, the map $F : \mathcal{S} \to \mathbb{R}$ is continuous.

Since $\mathfrak{P}_t$ is non-empty, compact-valued, and continuous (see Assumption 3.1 (ii)) and the map F is continuous, an application of Berge's maximum theorem (see, e.g., [2, Theorem 17.31]) ensures the continuity of $\widehat{J}_t$ and the existence of the measurable selector $\widehat{p}_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \widehat{p}_t(\cdot|s_t, a_t, \mu_{t:T}) \in \mathfrak{P}_t(s_t, a_t, \mu_t)$ satisfying (3.6).

Now let us prove (ii). To that end, denote by $\tilde{\mathbb{P}} := \widehat{p}_t(\cdot|s_t, a_t, \tilde{\mu}_{t:T}) \in \mathfrak{P}_t(s_t, a_t, \tilde{\mu}_t)$ where $\widehat{p}_t$ denotes the measurable selector given in Lemma 5.1 (i). Furthermore, by Assumption 3.1 (ii), we can choose $\mathbb{P} \in \mathfrak{P}_t(s_t, a_t, \mu_t)$ such that the following hold:

$$d_{W_1}(\mathbb{P}, \tilde{\mathbb{P}}) \leq L_{\mathfrak{P}_t} d_{W_1}(\mu_t, \tilde{\mu}_t), \tag{5.2}$$

and

$$\begin{aligned}\widehat{J}_t(s_t,a_t,\mu_{t:T})-\widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\leq&\int_S\Big(r(s_t,a_t,s_{t+1},\mu_t)+\widehat{V}_{t+1}(s_{t+1},\mu_{t+1:T})\Big)\,\mathbb{P}(ds_{t+1})\\&-\int_S\Big(r(s_t,a_t,\tilde{s}_{t+1},\tilde{\mu}_t)+\widehat{V}_{t+1}(\tilde{s}_{t+1},\tilde{\mu}_{t+1:T})\Big)\,\tilde{\mathbb{P}}(d\tilde{s}_{t+1})\\&=:\mathrm{B}(\mathbb{P},\tilde{\mathbb{P}}).\end{aligned}$$

Furthermore, since for every[7] $\gamma\in\mathrm{Cpl}(\mathbb{P},\tilde{\mathbb{P}})$, by Assumption 3.1 (i), (iii), and (5.1), we have

$$\begin{aligned}\mathrm{B}(\mathbb{P},\tilde{\mathbb{P}})=&\int_{S\times S}\Big(r(s_t,a_t,s_{t+1},\mu_t)-r(s_t,a_t,\tilde{s}_{t+1},\mu_t)+r(s_t,a_t,\tilde{s}_{t+1},\mu_t)-r(s_t,a_t,\tilde{s}_{t+1},\tilde{\mu}_t)\\&+\widehat{V}_{t+1}(s_{t+1},\mu_{t+1:T})-\widehat{V}_{t+1}(\tilde{s}_{t+1},\mu_{t+1:T})\\&+\widehat{V}_{t+1}(\tilde{s}_{t+1},\mu_{t+1:T})-\widehat{V}_{t+1}(\tilde{s}_{t+1},\tilde{\mu}_{t+1:T})\Big)\,\gamma(ds_{t+1},d\tilde{s}_{t+1})\\\leq&\int_{S\times S}\Big(L_r'|s_{t+1}-\tilde{s}_{t+1}|+L_r d_{W_1}(\mu_t,\tilde{\mu}_t)\\&+\widehat{L}_{t+1}'|s_{t+1}-\tilde{s}_{t+1}|+\widehat{L}_{t+1}\sum_{u=t+1}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)\Big)\gamma(ds_{t+1},d\tilde{s}_{t+1}),\end{aligned}$$

where $L_r',\widehat{L}_{t+1}'>0$ can be chosen appropriately thanks to Assumption 3.1 (i).

It thus holds that

$$\begin{aligned}\mathrm{B}(\mathbb{P},\tilde{\mathbb{P}})&\leq\widehat{K}_t\bigg(\sum_{u=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)+\inf_{\gamma\in\mathrm{Cpl}(\mathbb{P},\tilde{\mathbb{P}})}\int_{S\times S}|s_{t+1}-\tilde{s}_{t+1}|\gamma(ds_{t+1},d\tilde{s}_{t+1})\bigg)\\&=\widehat{K}_t\bigg(\sum_{u=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)+d_{W_1}(\mathbb{P},\tilde{\mathbb{P}})\bigg).\end{aligned}$$

where $\widehat{K}_t:=(L_r+L_r'+\widehat{L}_{t+1}+\widehat{L}_{t+1}')>0$.

Combined with (5.2), this ensure that

$$\widehat{J}_t(s_t,a_t,\mu_{t:T})-\widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\leq\widehat{K}_t(1+L_{\mathfrak{P}_t})\sum_{u=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u).$$

Using the same arguments as those used in the above upper bound, we can obtain the lower bound $\widehat{J}_t(s_t,a_t,\mu_{t:T})-\widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\geq-\widehat{K}_t(1+L_{\mathfrak{P}_t})\sum_{u=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)$, by using the same constant $\widehat{K}_t>0$. This completes the proof.

The proof of part (iii) follows from similar arguments as those used in the proof of (i). We define a map $G:S\times(\mathcal{P}(S))^{T-t}\times\mathcal{P}(A)\ni(s_t,\mu_{t:T},\pi_t)\to G(s_t,\mu_{t:T},\pi_t)\in\mathbb{R}$ by

$$G(s_t,\mu_{t:T},\pi_t):=\int_A\widehat{J}_t(s_t,a_t,\mu_{t:T})\pi(da_t).$$

Then we consider a sequence $(s_t^n,\mu_{t:T}^n,\pi_t^n)_{n\in\mathbb{N}}\subseteq S\times(\mathcal{P}(S))^{T-t}\times\mathcal{P}(A)$ such that $s_t^n\to s_t^\star$ $\mu_u^n\rightharpoonup\mu_u^\star$ (for every $s=t,\dots,T-1$), and $\pi_t^n\rightharpoonup\pi_t^\star$, as $n\to\infty$ with some $(s_t^\star,\mu_{t:T}^\star,\pi_t^\star)\in S\times\mathcal{P}(S)^{(T-t)}\times\mathcal{P}(A)$.

By the triangle inequality, for every $n\in\mathbb{N}$,

$$|G(s_t^n,\mu_{t:T}^n,\pi_t^n)-G(s_t^\star,\mu_{t:T}^\star,\pi_t^\star)|$$

[7] We refer to Section 2.1 for the definition of $\mathrm{Cpl}(\mathbb{P},\tilde{\mathbb{P}})$.

$$\begin{aligned}&\leq |G(s_t^\star,\mu_{t:T}^\star,\pi_t^n)-G(s_t^\star,\mu_{t:T}^\star,\pi_t^\star)|+|G(s_t^n,\mu_{t:T}^n,\pi_t^n)-G(s_t^\star,\mu_{t:T}^\star,\pi_t^n)|\\ &=: \mathrm{III}^n+\mathrm{IV}^n\,.\end{aligned}$$

We will show that III^n and IV^n vanish as $n\to\infty$.

Since $\widehat{J}_t(s_t^\star,\cdot,\mu_{0:T-t}^\star)$ is continuous on A (see Lemma 5.1 (i)) and the action space A is finite (see Assumption 3.1 (i)), the limit $\pi_t^n \rightharpoonup \pi_t^\star$ ensures that III^n vanishes as $n\to\infty$.

Furthermore, as S is also finite (see Assumption 3.1 (i)), there exists $N\in\mathbb{N}$ such that for every $n\geq N$ we have $s_t^n=s_t^\star$. By Lemma 5.1 (ii), we then have for every $n\geq N$,

$$\mathrm{IV}^n\leq\int_A\left|\widehat{J}_t(s_t^\star,a_t,\mu_{t:T}^n)-\widehat{J}_t(s_t^\star,a_t,\mu_{t:T}^\star)\right|\pi_t^n(da_t)\leq\widehat{K}_t\sum_{u=t}^{T-1}d_{W_1}(\mu_u^n,\mu_u^\star).$$

Combined with the limit $\mu_i^n\rightharpoonup\mu_u^\star$ (for every $u=t,\dots,T-1$), this ensures that IV^n vanish as $n\to\infty$. Therefore, the map G is continuous.

Since $\mathcal{P}(A)$ is compact (noting that A is finite) and G is continuous, an application of Berge's maximum theorem ensures the continuity of $\widehat{V}_t$ and the existence of the measurable selector $\widehat{\pi}_t: S\times(\mathcal{P}(S))^{T-t}\ni(s_t,\mu_{t:T})\mapsto\widehat{\pi}_t(\cdot|s_t,\mu_{t:T})\in\mathcal{P}(A)$ satisfying (3.7).

Lastly we prove the part (iv). By Assumption 3.1 (i), (iii), and (5.1),

$$\begin{aligned}|\widehat{V}_t(s_t,\mu_{t:T})|&\leq\sup_{\pi\in\mathcal{P}(A)}\int_A\inf_{\mathbb{P}\in\mathfrak{P}_t(s_t,a_t,\mu_t)}\int_X\left(|r(s_t,a_t,s_{t+1},\mu_t)|+|\widehat{V}_{t+1}(s_{t+1},\mu_{t+1:T})|\right)\mathbb{P}(dy)\pi(da)\\&\leq C_r+\widehat{C}_{t+1}.\end{aligned}$$

By letting $\widehat{C}_t:=C_r+\widehat{C}_{t+1}$, we have $|\widehat{V}_t(s_t,\mu_{t:T})|\leq\widehat{C}_t$.

To have the other estimates, denote by $\pi:=\widehat{\pi}_t(\cdot|s_t,\mu_{t:T})\in\mathcal{P}(A)$ where $\widehat{\pi}_t$ is the measurable selector given in Lemma 5.1 (iii). Then since π is not necessarily a maximizer for $\widehat{V}_t(s_t,\tilde{\mu}_{t:T})$ but for $\widehat{V}_t(s_t,\mu_{t:T})$, it holds

$$\widehat{V}_t(s_t,\mu_{t:T})-\widehat{V}_t(s_t,\tilde{\mu}_{t:T})\leq\int_A\left(\widehat{J}_t(s_t,a_t,\mu_{t:T})-\widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\right)\pi(da_t).\tag{5.3}$$

Further, by Lemma 5.1 (ii), $\int_A\widehat{J}_t(s_t,a_t,\mu_{t:T})-\widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\pi(da_t)\leq\widehat{K}_t\sum_{u=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)$, which leads to the upper bound estimates with letting $\widehat{L}_t:=\widehat{K}_t$.

Using the same arguments as those used in the above estimates, we can have $\widehat{V}_t(s_t,\mu_{t:T})-\widehat{V}_t(s_t,\tilde{\mu}_{t:T})\geq-\widehat{L}_t\sum_{s=t}^{T-1}d_{W_1}(\mu_u,\tilde{\mu}_u)$, with the same constant $\widehat{L}_t>0$. This completes the proof. □

Proof of Lemma 3.2. We will prove the parts (i) and (ii) together. First we claim that when $t=T-1$, there exists a measurable selector $\widehat{p}_{T-1}: S\times A\times\mathcal{P}(S)\ni(s_{T-1},a_{T-1},\mu_{T-1})\mapsto\widehat{p}_{T-1}(\cdot|s_{T-1},a_{T-1},\mu_{T-1})\in\mathfrak{P}_{T-1}(s_{T-1},a_{T-1},\mu_{T-1})$ satisfying (3.5). Indeed, since $\widehat{J}_{T-1}$ has a simple integrand $r(\cdot,\cdot,\cdot,\cdot)$ (see (3.2)), the same arguments as for the proof of Lemma 5.1 (i) (applying Berge's maximum theorem), but with respect to the map $F:\mathcal{S}\to\mathbb{R}$ given by

$$F(s_{T-1},a_{T-1},\mu_{T-1},p_{T-1}):=\int_S r(s_{T-1},a_{T-1},s_T,\mu_{T-1})p_{T-1}(ds_T),$$

with $\mathcal{S}=\{(s_{T-1},a_{T-1},\mu_{T-1})\in S\times A\times\mathcal{P}(S),\ p_{T-1}\in\mathfrak{P}_{T-1}(s_{T-1},a_{T-1},\mu_{T-1})\}$, ensure the existence of the selector $\widehat{p}_{T-1}$.

Analogously, when $t=T-1$, there exists a measurable selector $\widehat{\pi}_{T-1}: S\times\mathcal{P}(S)\ni(s_{T-1},\mu_{T-1})\mapsto\widehat{\pi}_{T-1}(\cdot|s_{T-1},\mu_{T-1})\in\mathcal{P}(A)$ satisfying (3.7). Indeed, we first claim that there is $\widehat{K}_{T-1}>0$ such that for every $s_{T-1} \in S$, $a_{T-1} \in A$, $\mu_{T-1}, \tilde{\mu}_{T-1} \in \mathcal{P}(S)$, it holds that

$$|\widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \mu_{T-1}) - \widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \tilde{\mu}_{T-1})| \leq \widehat{K}_{T-1} d_{W_1}(\mu_{T-1}, \tilde{\mu}_{T-1}). \tag{5.4}$$

By the existence of $\widehat{p}_{T-1}$ satisfying (3.5), the arguments devoted for the proof of Lemma 5.1 (ii) using $\widehat{p}_{T-1}$ and Assumptions 3.1 (i), (iii) ensure that we have $\widehat{K}_{T-1} > 0$ satisfying (5.4).

By (5.4), we can use the same arguments presented for the proof of Lemma 5.1 (iii) using Berge's maximum theorem to have the existence of the measurable selector $\widehat{\pi}_{T-1}$ satisfying (3.7).

So far we have proven (i) and (ii) for the case $t = T-1$. The other cases (i.e., $t \leq T-2$) can be proven by applying Lemma 5.1 under the condition of the existence of constants $\widehat{C}_{T-1} \geq 1$, $\widehat{L}_{T-1} > 0$ such that for every $s_{T-1} \in S$ and $\mu_{T-1}, \tilde{\mu}_{T-1} \in \mathcal{P}(S)$, it holds

$$\begin{aligned} |\widehat{V}_{T-1}(s_{T-1}, \mu_{T-1})| &\leq \widehat{C}_{T-1}, \\ \left|\widehat{V}_{T-1}(s_{T-1}, \mu_{T-1}) - \widehat{V}_{T-1}(s_{T-1}, \tilde{\mu}_{T-1})\right| &\leq \widehat{L}_{T-1} d_{W_1}(\mu_{T-1}, \tilde{\mu}_{T-1}). \end{aligned}$$

By the existence of $\widehat{p}_{T-1}$ and $\widehat{\pi}_{T-1}$ and the estimates given in (5.4), we can use the same arguments presented for the proof of Lemma 5.1 (iv) to obtain those constants satisfying the above estimates. □

Proof of Proposition 3.4. By the existence of $\widehat{p}_{0:T}$ and $\widehat{\pi}_{0:T}$ given in Lemma 3.2, it is straightforward to prove the part (i). Indeed for every $t = 0, \ldots, T-1$, we can define sequences of stochastic kernels by for every $(s_t, a_t, \mu_t) \in S \times A \times \mathcal{P}(S)$,

$$p_t^*(\cdot|s_t, a_t, \mu_t) := \begin{cases} \widehat{p}_t(\cdot|s_t, a_t, \mu_t, \tilde{\mu}_{t+1:T}) & \text{if } t \leq T-2; \\ \widehat{p}_t(\cdot|s_t, a_t, \mu_t) & \text{if } t = T-1, \end{cases}$$

and for every $s_t \in S$,

$$\pi_t^*(\cdot|s_t) := \widehat{\pi}_t(\cdot|s_t, \tilde{\mu}_{t:T}).$$

By the optimality of $\widehat{p}_{0:T}$ and $\widehat{\pi}_{0:T}$ (see (3.5)-(3.7)), $p^*_{0:T}$ and $\pi^*_{0:T}$ constructed above satisfy (3.8)-(3.10).

Now let us prove (ii). Let $\overline{\mathbb{P}} := \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi_0^*, p_0)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi^*_{T-1}, p_{T-1})} \in \mathcal{Q}(\tilde{\mu}_{0:T}, \pi^*_{0:T})$ and denote by for every $t = 1, \ldots, T-1$, $\overline{\mathbb{P}}_{0:t} := \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi_0^*, p_0)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_t, \pi_t^*, p_t)}$ and $\overline{\mathbb{P}}_0 = \tilde{\mu}_0$.

Note that by the definitions of $\widehat{V}_{0:T}$ and $\widehat{J}_{0:T}$ given in (3.1)-(3.3) and the optimality of $\pi^*_{0:T}$ given in (3.10),

$$\begin{aligned} &\mathbb{E}^{\overline{\mathbb{P}}}\Big[r(s_{T-1}, a_{T-1}, s_T, \tilde{\mu}_{T-1})\Big] \\ &= \mathbb{E}^{\overline{\mathbb{P}}_{0:T-1}}\left[\int_{S \times A} r(s_{T-1}, a_{T-1}, s_T, \tilde{\mu}_{T-1}) \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi^*_{T-1}, p_{T-1})}(ds_T, da_{T-1}|s_{T-1})\right] \\ &\geq \mathbb{E}^{\overline{\mathbb{P}}_{0:T-1}}\left[\int_A \widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \tilde{\mu}_{T-1}) \pi^*_{T-1}(da_{T-1}|s_{T-1})\right] = \mathbb{E}^{\overline{\mathbb{P}}}\Big[\widehat{V}_{T-1}(s_{T-1}, \tilde{\mu}_{T-1})\Big], \end{aligned} \tag{5.5}$$

and that for every $t \leq T-2$,

$$\begin{aligned} &\mathbb{E}^{\overline{\mathbb{P}}}\Big[r(s_t, a_t, s_{t+1}, \tilde{\mu}_t) + \widehat{V}_{t+1}(s_{t+1}, \tilde{\mu}_{t+1:T})\Big] \\ &= \mathbb{E}^{\overline{\mathbb{P}}_{0:t}}\left[\int_{S \times A} \Big(r(s_t, a_t, s_{t+1}, \tilde{\mu}_t) + \widehat{V}_{t+1}(s_{t+1}, \tilde{\mu}_{t+1:T})\Big) \mathbb{P}_{(\tilde{\mu}_t, \pi_t^*, p_t)}(ds_{t+1}, da_t|s_t)\right] \\ &\geq \mathbb{E}^{\overline{\mathbb{P}}_{0:t}}\left[\int_A \widehat{J}_t(s_t, a_t, \tilde{\mu}_{t:T}) \pi_t^*(da_t|s_t)\right] = \mathbb{E}^{\overline{\mathbb{P}}}\Big[\widehat{V}_t(s_t, \tilde{\mu}_{t:T})\Big]. \end{aligned} \tag{5.6}$$

By (5.5) and (5.6), we hence have

$$\begin{aligned}\mathbb{E}^{\overline{\mathbb{P}}}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right] &= \mathbb{E}^{\overline{\mathbb{P}}}\left[\sum_{t=0}^{T-2} r(s_t,a_t,s_{t+1},\tilde{\mu}_t) + r(s_{T-1},a_{T-1},s_T,\tilde{\mu}_{T-1})\right]\\ &\geq \mathbb{E}^{\overline{\mathbb{P}}}\left[\sum_{t=0}^{T-2} r(s_t,a_t,s_{t+1},\tilde{\mu}_t) + \widehat{V}_{T-1}(s_{T-1},\tilde{\mu}_{T-1})\right]\\ &\geq \cdots \geq \mathbb{E}^{\overline{\mathbb{P}}}\Big[\widehat{V}_0(s_0,\tilde{\mu}_{0:T})\Big] = \widehat{V}(\tilde{\mu}_{0:T}).\end{aligned}$$

Since $\overline{\mathbb{P}}$ is arbitrary in $\mathcal{Q}(\tilde{\mu}_{0:T},\pi^*_{0:T})$, we have

$$\inf_{\overline{\mathbb{P}}\in\mathcal{Q}(\tilde{\mu}_{0:T},\pi^*_{0:T})} \mathbb{E}^{\overline{\mathbb{P}}}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right] \geq \widehat{V}(\tilde{\mu}_{0:T}) = \mathbb{E}^{\mathbb{P}^*(\tilde{\mu}_{0:T})}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right],$$

with $\mathbb{P}^*(\tilde{\mu}_{0:T}) := \mathbb{P}(\tilde{\mu}_{0:T},\pi^*_{0:T},p^*_{0:T}) = \tilde{\mu}_0\otimes\mathbb{P}_{(\tilde{\mu}_0,\pi^*_0,p^*_0)}\otimes\cdots\otimes\mathbb{P}_{(\tilde{\mu}_{T-1},\pi^*_{T-1},p^*_{T-1})}\in\mathcal{Q}(\tilde{\mu}_{0:T},\pi^*_{0:T})$. Furthermore, since $\pi^*_{0:T}\in\Pi$, we hence have $V(\tilde{\mu}_{0:T})\geq\widehat{V}(\tilde{\mu}_{0:T})$.

Let $\pi_{0:T}\in\Pi$ and $\underline{\mathbb{P}} := \tilde{\mu}_0\otimes\mathbb{P}_{(\tilde{\mu}_0,\pi_0,p^*_0)}\otimes\cdots\otimes\mathbb{P}_{(\tilde{\mu}_{T-1},\pi_{T-1},p^*_{T-1})}\in\mathcal{Q}(\tilde{\mu}_{0:T},\pi_{0:T})$ and denote by for every $t=1,\dots,T-1$, $\underline{\mathbb{P}}_{0:T} := \tilde{\mu}_0\otimes\mathbb{P}_{(\tilde{\mu}_0,\pi_0,p^*_0)}\otimes\cdots\otimes\mathbb{P}_{(\tilde{\mu}_t,\pi_t,p^*_t)}$ and $\underline{\mathbb{P}}_0=\tilde{\mu}_0$.

From the definitions of $\widehat{V}_{0:T}$ and $\widehat{J}_{0:T}$ given in (3.1)-(3.3) and the optimality of $p^*_{0:T}$ given in (3.8) and (3.9), it follows that

$$\begin{aligned}&\mathbb{E}^{\underline{\mathbb{P}}}\Big[r(s_{T-1},a_{T-1},s_T,\tilde{\mu}_{T-1})\Big]\\ &= \mathbb{E}^{\underline{\mathbb{P}}_{0:T-1}}\left[\int_{S\times A} r(s_{T-1},a_{T-1},s_T,\tilde{\mu}_{T-1})\mathbb{P}_{(\tilde{\mu}_{T-1},\pi_{T-1},p^*_{T-1})}(ds_T,da_{T-1}|s_{T-1})\right]\\ &= \mathbb{E}^{\underline{\mathbb{P}}_{0:T-1}}\left[\int_A \widehat{J}_{T-1}(s_{T-1},a_{T-1},\tilde{\mu}_{T-1})\pi_{T-1}(da_{T-1}|x_{T-1})\right] \leq \mathbb{E}^{\underline{\mathbb{P}}}\Big[\widehat{V}_{T-1}(s_{T-1},\tilde{\mu}_{T-1})\Big],\end{aligned}\tag{5.7}$$

and that for every $t\leq T-2$,

$$\begin{aligned}&\mathbb{E}^{\underline{\mathbb{P}}}\Big[r(s_t,a_t,s_{t+1},\tilde{\mu}_t)+\widehat{V}_{t+1}(s_{t+1},\tilde{\mu}_{t+1:T})\Big]\\ &= \mathbb{E}^{\underline{\mathbb{P}}_{0:t}}\left[\int_{S\times A}\Big(r(s_t,a_t,s_{t+1},\tilde{\mu}_t)+\widehat{V}_{t+1}(s_{t+1},\tilde{\mu}_{t+1:T})\Big)\,\mathbb{P}_{(\tilde{\mu}_t,\pi_t,p^*_t)}(ds_{t+1},da_t|s_t)\right]\\ &= \mathbb{E}^{\underline{\mathbb{P}}_{0:t}}\left[\int_A \widehat{J}_t(s_t,a_t,\tilde{\mu}_{t:T})\pi_t(da_t|s_t)\right] \leq \mathbb{E}^{\underline{\mathbb{P}}}\Big[\widehat{V}_t(s_t,\tilde{\mu}_{t:T})\Big].\end{aligned}\tag{5.8}$$

By (5.7) and (5.8), we hence have

$$\begin{aligned}\mathbb{E}^{\underline{\mathbb{P}}}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right] &= \mathbb{E}^{\underline{\mathbb{P}}}\left[\sum_{t=0}^{T-2} r(s_t,a_t,s_{t+1},\tilde{\mu}_t) + r(s_{T-1},a_{T-1},s_T,\tilde{\mu}_{T-1})\right]\\ &\leq \mathbb{E}^{\underline{\mathbb{P}}}\left[\sum_{t=0}^{T-2} r(s_t,a_t,s_{t+1},\tilde{\mu}_t) + \widehat{V}_{T-1}(s_{T-1},\tilde{\mu}_{T-1})\right]\\ &\leq \cdots \leq \mathbb{E}^{\underline{\mathbb{P}}}\Big[\widehat{V}_0(s_0,\tilde{\mu}_{0:T})\Big] = \widehat{V}(\tilde{\mu}_{0:T}),\end{aligned}$$

which ensures that

$$\inf_{\mathbb{P}\in\mathcal{Q}(\tilde{\mu}_{0:T},\pi_{0:T})} \mathbb{E}^{\mathbb{P}}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right] \leq \mathbb{E}^{\underline{\mathbb{P}}}\left[\sum_{t=0}^{T-1} r(s_t,a_t,s_{t+1},\tilde{\mu}_t)\right] \leq \widehat{V}(\tilde{\mu}_{0:T}).\tag{5.9}$$

Since $\pi_{0:T}$ is arbitrary in Π, we have $V(\tilde{\mu}_{0:T})\leq\widehat{V}(\tilde{\mu}_{0:T})$.

It remains to show the equality of $V(\tilde{\mu}_{0:T})$ to the supremum in (3.11). Since the last inequality given in (5.9) holds for any $\pi_{0:T} \in \Pi$ (with recalling $\underline{\mathbb{P}} = \mathbb{P}(\tilde{\mu}_{0:T}, \pi_{0:T}, p^*_{0:T}) = \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi_0, p^*_0)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi_{T-1}, p^*_{T-1})} \in \mathcal{Q}(\tilde{\mu}_{0:T}, \pi_{0:T})$), it follows that

$$\sup_{\pi \in \Pi} \mathbb{E}^{\mathbb{P}(\tilde{\mu}_{0:T}, \pi_{0:T}, p^*_{0:T})}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \tilde{\mu}_t)\right] \leq \mathbb{E}^{\mathbb{P}^*(\tilde{\mu}_{0:T})}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \tilde{\mu}_t)\right] = V(\tilde{\mu}_{0:T})$$

where the last equality follows from above (i.e., $V(\tilde{\mu}_{0:T}) = \widehat{V}(\tilde{\mu}_{0:T})$).

On the other hand, since $\pi^*_{0:T} \in \Pi$ and $\mathbb{P}^*(\tilde{\mu}_{0:T}) = \mathbb{P}(\tilde{\mu}_{0:T}, \pi^*_{0:T}, p^*_{0:T}) = \tilde{\mu}_0 \otimes \mathbb{P}_{(\tilde{\mu}_0, \pi^*_0, p^*_0)} \otimes \cdots \otimes \mathbb{P}_{(\tilde{\mu}_{T-1}, \pi^*_{T-1}, p^*_{T-1})} \in \mathcal{Q}(\tilde{\mu}_{0:T}, \pi^*_{0:T})$, the above inequality establishes equality. This completes the proof. □

5.2. **Proof of Lemma 3.5 and Proposition 3.6.**

Proof of Lemma 3.5. Fix $\overline{\pi}^N_{0:T} \in \Pi^N$, let $t \leq T-1$ and set

$$\mathcal{S} := \left\{ (\overline{s}^N_t, \overline{a}^N_t, \overline{p}^N_t) \,\middle|\, (\overline{s}^N_t, \overline{a}^N_t) \in S^N \times A^N, \overline{p}^N_t \in \mathfrak{P}^N_t(\overline{s}^N_t, \overline{a}^N_t) \right\}.$$

Define an auxiliary map $F : \mathcal{S} \ni (\overline{s}^N_t, \overline{a}^N_t, \overline{p}^N_t) \mapsto F(\overline{s}^N_t, \overline{a}^N_t, \overline{p}^N_t) \in \mathbb{R}$ by

$$F(\overline{s}^N_t, \overline{a}^N_t, \overline{p}^N_t) := \int_{S^N} f(\overline{s}^N_t, \overline{a}^N_t, \overline{s}^N_{t+1}) \overline{p}^N_t(d\overline{s}^N_{t+1}),$$

where if $t = T-1$, then we set

$$f(\overline{s}^N_t, \overline{a}^N_t, \overline{s}^N_{t+1}) := r\big(s^i_t, a^i_t, s^i_{t+1}, e^N(\overline{s}^N_t)\big),$$

whereas if $t \leq T-2$, then we set

$$f(\overline{s}^N_t, \overline{a}^N_t, \overline{s}^N_{t+1}) := r\big(s^i_t, a^i_t, s^i_{t+1}, e^N(\overline{s}^N_t)\big) + \int_{A^N} \widehat{J}^N_{t+1,i}(\overline{s}^N_{t+1}, \overline{a}^N_{t+1}; \overline{\pi}^N_{0:T}) \overline{\pi}^N_{t+1}(d\overline{a}^N_{t+1} | \overline{s}^N_{t+1}).$$

Since both S^N and A^N are finite, F is continuous in $(\overline{s}^N_t, \overline{a}^N_t)$. Again, by the finiteness of S^N and A^N, we get that f is continuous. Hence, F is continuous in $\overline{p}^N_t$. From here, we can follow the same ideas as presented in the proofs of Lemma 3.2 and Lemma 5.1 to prove the result. □

Proof of Proposition 3.6. We can use the same approach as presented in the proof of Proposition 3.4 (ii) to show that

$$\begin{aligned} J^N_i(\mu^o, \overline{\pi}^N_{0:T}) &= \inf_{\overline{\mathbb{P}}^N \in \mathcal{Q}^N(\mu^o, \overline{\pi}^N_{0:T})} \mathbb{E}^{\overline{\mathbb{P}}^N}\left[\sum_{t=0}^{T-1} r\big(s^i_t, a^i_t, s^i_{t+1}, e^N(\overline{s}^N_t)\big)\right] \\ &= \mathbb{E}^{\overline{\mathbb{P}}^N_i(\mu^o, \overline{\pi}^N_{0:T}, \widehat{p}_{(0:T, i, \overline{\pi}^N_{0:T})})}\left[\sum_{t=0}^{T-1} r\big(s^i_t, a^i_t, s^i_{t+1}, e^N(\overline{s}^N_t)\big)\right], \end{aligned}$$

where the second equality follows by definition of $\widehat{J}^N_{0,i}$. □

6. Proof of results in Section 3.2

6.1. **Preliminary lemmas.**

Let us provide some simple observations that play an instrumental role in the proof of Proposition 3.9 and Theorem 3.10.

Let us begin with a measurable extension of mappings into stochastic kernels defined on probability spaces. The proof can be found in Appendix A.

Lemma 6.1. *Suppose that Assumption 3.1 is satisfied. Let $t \in \{0, \dots, T-1\}$ and $\tilde{\mu}_{t:T} \in (\mathcal{P}(S))^{T-t}$. Furthermore, let $p_t \colon S \times A \ni (s_t, a_t) \mapsto p_t(\cdot|s_t, a_t) \in \mathcal{P}(S)$ be a mapping. Then there exists a Borel-measurable mapping (i.e., stochastic kernel) $\overline{p}_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \overline{p}_t(\cdot|s_t, a_t, \mu_{t:T}) \in \mathcal{P}(S)$ such that for every $(s_t, a_t) \in S \times A$*

$$\overline{p}_t(\cdot|s_t, a_t, \hat{\mu}_{t:T}) = p_t(\cdot|s_t, a_t).$$

The following two lemmas link the correspondences $\mathcal{C}, \mathcal{B}$ (given in Definition 3.8 (i)) into the dynamic programming results given in Lemma 3.2 (and Proposition 3.4).

Lemma 6.2. *Suppose that Assumption 3.1 is satisfied. Let $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$ and denote by $p^{\tilde{\nu}}_{0:T-1}$ the corresponding kernels enabling $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$ (see Definition 3.8 (i)). For every $t = 0, \dots, T-1$, define $\overline{p}^{\tilde{\nu}}_t : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \mapsto \overline{p}^{\tilde{\nu}}_t(\cdot|s_t, a_t, \mu_t) \in \mathcal{P}(S)$ by*

$$\overline{p}^{\tilde{\nu}}_t(\cdot|s_t, a_t, \mu_t) := \begin{cases} p^{\tilde{\nu}}_t(\cdot|s_t, a_t, \mu_t, \nu_{t+1:T,S}) & \textit{if } t \leq T-2; \\ \widehat{p}_t(\cdot|s_t, a_t, \mu_t) & \textit{if } t = T-1, \end{cases}$$

where $\widehat{p}_t$ is the stochastic kernel given in Lemma 3.2 (i). Then for every $(s_t, a_t) \in S \times A$, $\overline{p}^{\tilde{\nu}}_t(\cdot|s_t, a_t, \nu_{t,S})$ is optimal for $\widehat{J}_t(s_t, a_t, \nu_{t:T,S})$, i.e., if $t = T-1$,

$$\int_S r(s_{T-1}, a_{T-1}, s_T, \nu_{T-1,S}) \overline{p}^{\tilde{\nu}}_{T-1}(ds_T|s_{T-1}, a_{T-1}, \nu_{T-1,S}) = \widehat{J}_{T-1}(s_{T-1}, a_{T-1}, \nu_{T-1,S}),$$

whereas if $t \leq T-2$,

$$\int_S \Big(r(s_t, a_t, s_{t+1}, \nu_{t,S}) + \widehat{V}_{t+1}(s_{t+1}, \nu_{t+1:T,S}) \Big) \overline{p}^{\tilde{\nu}}_t(ds_{t+1}|s_t, a_t, \nu_{t,S}) = \widehat{J}_t(s_t, a_t, \nu_{t:T,S}).$$

Proof. It is straightforward to show the case where $t = T-1$ by the optimality of $\widehat{p}_{T-1} (= p^{\tilde{\nu}}_{T-1})$ presented in Lemma 3.2 (i). For the case where $t \leq T-2$, since for every $(s_t, a_t) \in S \times A$

$$\overline{p}^{\tilde{\nu}}_t(\cdot|s_t, a_t, \nu_{t,S}) = p^{\hat{\nu}}_t(\cdot|s_t, a_t, \nu_{t:T,S}) \in \widetilde{\mathfrak{P}}_t(s_t, a_t, \nu_{t:T,S})$$

(see Definition 3.8 (i) and Remark 3.3), $\overline{p}^{\tilde{\nu}}_t(\cdot|s_t, a_t, \nu_{t,S})$ is optimal for $\widehat{J}_t(s_t, a_t, \nu_{t:T,S})$. This completes the proof. □

Lemma 6.3. *Suppose that Assumption 3.1 is satisfied. Let $\nu_{0:T}$, $\tilde{\nu}_{0:T} \in \Xi$ and denote by $\pi^{\tilde{\nu}}_{0:T}$ the disintegrating kernels of $\tilde{\nu}_{0:T}$ (see Definition 3.7). Furthermore, denote for every $t = 0, \dots, T-1$ by $\tilde{w}_t(s)$ the weight of the measure $\tilde{\nu}_{t,S}(\cdot)$ at each point $s \in S$ (i.e., $\sum_{s \in S} \tilde{w}_t(s) = 1$ with $\tilde{w}_t(s) \geq 0$ for $s \in S$). Then the following hold:*

(i) $\tilde{\nu}_{0:T} \in \mathcal{B}(\nu_{0:T})$ (see Definition 3.8 (i)) if and only if for every $t = 0, \dots, T-1$ and $s_t \in S$ such that $\tilde{w}_t(s_t) > 0$, $\pi^{\tilde{\nu}}_t(\cdot|s_t) \in \mathcal{P}(A)$ is optimal for $\widehat{V}_t(s_t, \nu_{t:T,S})$ (see (3.1)).

(ii) Let $\tilde{\nu}_{0:T} \in \mathcal{B}(\nu_{0:T})$. For every $t = 0, \dots, T-1$, define $\overline{\pi}^{\tilde{\nu}}_t : S \ni s_t \mapsto \overline{\pi}^{\tilde{\nu}}_t(\cdot|s_t) \in \mathcal{P}(A)$ by

$$\overline{\pi}^{\tilde{\nu}}_t(\cdot|s_t) := \begin{cases} \pi^{\hat{\nu}}_t(\cdot|s_t) & \textit{if } \tilde{w}_t(s_t) > 0; \\ \widehat{\pi}_t(\cdot|s_t, \nu_{t:T,S}) & \textit{else}, \end{cases} \tag{6.1}$$

where $\widehat{\pi}_t$ is the measurable selector given in Lemma 3.2. Then it holds

$$\tilde{\nu}_t(ds_t, da_t) = \overline{\pi}^{\tilde{\nu}}_t(da_t|s_t) \tilde{\nu}_{t,S}(ds_t). \tag{6.2}$$

Furthermore, $\overline{\pi}^{\tilde{\nu}}_t(\cdot|s_t)$ is optimal for $\widehat{V}_t(s_t, \nu_{t:T,S})$ for every $s_t \in S$.

Proof. We start by proving the statement (i). Suppose $\tilde{\nu}_{0:T} \in \mathcal{B}(\nu_{0:T})$. Fix any $t = 0, \dots, T-1$. Then since $\tilde{\nu}_t(D_t(\nu_{t:T})) = 1$,

$$1 = \int_S \int_A \mathbf{1}_{\{(s_t, a_t) \in D_t(\nu_{t:T})\}} \pi^{\hat{\nu}}_t(da_t|s_t) \tilde{\nu}_{t,S}(ds_t)$$

$$= \sum_{s_t \in S} \tilde{w}_t(s_t) \pi_t^{\tilde{\nu}} \big(\{ a_t \in A | (s_t, a_t) \in D_t(\nu_{t:T}) \} \big| s_t \big).$$

This implies that for every $s_t \in S$ such that $\tilde{w}_t(s_t) > 0$, $\pi_t^{\tilde{\nu}} \big(\{ a_t \in A | (s_t, a_t) \in D_t(\nu_{t:T}) \} \big| s_t \big) = 1$.

We hence have that for every $s_t \in S$ such that $\tilde{w}_t(s_t) > 0$, it holds

$$\begin{aligned} \int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \pi_t^{\tilde{\nu}}(da_t | s_t) &= \int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \mathbf{1}_{\{a_t \in A | (s_t, a_t) \in D_t(\nu_{t:T})\}} \pi_t^{\tilde{\nu}}(da_t | s_t) \\ &= \int_A \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}) \mathbf{1}_{\{a_t \in A | (s_t, a_t) \in D_t(\nu_{t:T})\}} \pi_t^{\tilde{\nu}}(da_t | s_t) \\ &= \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}). \end{aligned} \tag{6.3}$$

Furthermore, since

$$\max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}) \geq \sup_{\pi \in \mathcal{P}(A)} \int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \pi(da_t) = \widehat{V}_t(s_t, \nu_{t:T,S}),$$

it follows from $\pi_t^{\tilde{\nu}}(da_t | s_t) \in \mathcal{P}(A)$ and (6.3) that $\pi_t^{\tilde{\nu}}(\cdot | s_t)$ is optimal for $\widehat{V}_t(s_t, \nu_{0:T,S})$.

Now suppose that for every $t = 0, \dots, T-1$ and $s_t \in S$ such that $\tilde{w}_t(s_t) > 0$, $\pi_t^{\tilde{\nu}}(\cdot | s_t)$ is optimal for $\widehat{V}_t(s_t, \nu_{0:T,S})$. Assume that there exists some $t \leq T-1$ such that $\tilde{\nu}_t(D_t(\nu_{t:T})) < 1$.

Set $S' := \big\{ s_t \in S \big| \pi_t^{\tilde{\nu}}(\{a_t \in A | (s_t, a_t) \in D_t(\nu_{t:T})\} | s_t) < 1 \text{ and } \tilde{w}_t(s_t) > 0 \big\}$, which is non-empty (due to $\tilde{\nu}_t(D_t(\nu_{t:T})) < 1$). Define for every $s_t \in S'$ by

$$A'(s_t) := \{ a_t \in A | (s_t, a_t) \notin D_t(\nu_{t:T}) \}.$$

Let $s_t \in S'$ and denote by $w_{t,s_t}(a_t)$ the weight of $\pi_t^{\tilde{\nu}}(\cdot | s_t)$ at $a_t \in A$. We now define $\pi_t' \in \mathcal{P}(A)$ by for every Borel set $E \in \mathcal{B}_A$,

$$\pi_t'(E) = \sum_{a_t \in A} \frac{w_{t,s_t}(a_t)}{1 - \sum_{a_t' \in A'(s_t)} w_{t,s_t}(a_t')} \mathbf{1}_{\{a_t \in E \setminus A'(s_t)\}}. \tag{6.4}$$

Then since $\pi_t^{\tilde{\nu}}(\{a_t \in A | \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) < \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S})\} | s_t) > 0$ (due to $s_t \in S'$),

$$\begin{aligned} \int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \pi_t^{\tilde{\nu}}(da_t | s_t) &< \int_A \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}) \pi_t^{\tilde{\nu}}(da_t | s_t) \\ &= \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}). \end{aligned} \tag{6.5}$$

Furthermore, since $\pi_t'(A'(s_t)) = \pi_t'\big(\{a_t \in A \big| \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) < \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S})\}\big) = 0$,

$$\begin{aligned} \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}) &= \int_A \max_{a_t' \in A} \widehat{J}_t(s_t, a_t', \nu_{t:T,S}) \pi_t'(da_t) \\ &= \int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \pi_t'(da_t) \leq \widehat{V}_t(s_t, \nu_{t:T,S}). \end{aligned}$$

Combining this with (6.5) implies that $\int_A \widehat{J}_t(s_t, a_t, \nu_{t:T,S}) \pi_t^{\tilde{\nu}}(da_t | s_t) < \widehat{V}_t(s_t, \nu_{t:T,S})$, which is a contradiction to the optimality of $\pi_t^{\tilde{\nu}}(\cdot | s_t)$ for $\widehat{V}_t(s_t, \nu_{0:T,S})$.

Thus, $\tilde{\nu}_t(D_t(\nu_{t:T})) = 1$ for every $t = 0, \dots, T-1$, i.e., $\tilde{\nu}_{0:T} \in \mathcal{B}(\nu_{0:T})$.

Now let us prove (ii). By the construction given in (6.1), it is straightforward to see that (6.2) holds. Hence it remains to show the optimality of $\overline{\pi}_t^{\tilde{\nu}}(\cdot | s_t)$ for $\widehat{V}_t(s_t, \nu_{t:T,S})$ for every $s_t \in S$.

Let $s_t \in S$ be such that $\tilde{w}_t(s_t) > 0$. Then Lemma 6.3 (i) ensures that $\overline{\pi}_t^{\tilde{\nu}}(\cdot | s_t) = \pi_t^{\tilde{\nu}}(\cdot | s_t)$ is optimal for $\widehat{V}_t(s_t, \mu_{t:T,S})$. For the other case where $s_t \in S$ with $\tilde{w}_t(s_t) = 0$, since $\overline{\pi}_t^{\tilde{\nu}}(\cdot | s_t) = \widehat{\pi}_t(\cdot | s_t, \nu_{t:T,S})$, the optimality of $\widehat{\pi}_t$ given in Lemma 3.2 (ii) ensures that $\overline{\pi}_t^{\tilde{\nu}}(\cdot | s_t)$ is optimal for $\widehat{V}_t(s_t, \mu_{t:T,S})$. This completes the proof. □

6.2. **Proof of Proposition 3.9.**

Proof of Proposition 3.9 (i). We first note that by the existence of $\widehat{p}_{0:T}$ given in Lemma 3.2 (i), $\mathcal{C}$ (given in Definition 3.8 (i)) is non-empty.

We claim that Γ is non-empty. To that end, let $\nu_{0:T} \in \Xi$ and choose an arbitrary $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$. Now for every $t = 0, \dots, T-1$, set

$$\tilde{\nu}'_t(ds_t, da_t) := \widehat{\pi}_t(da_t|s_t, \nu_{t:T,S})\tilde{\nu}_{t,S}(ds_t),$$

where $\widehat{\pi}_t$ is the measurable selector given in Lemma 3.2 (ii).

Then since $\tilde{\nu}'_{t,S}(\cdot) = \tilde{\nu}_{t,S}(\cdot)$ and $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$, it is clear that $\tilde{\nu}'_{0:T} \in \mathcal{C}(\nu_{0:T})$. Hence it remains to show that $\tilde{\nu}'_{0:T} \in \mathcal{B}(\nu_{0:T})$. Indeed, since the disintegrating kernel $\pi_t^{\tilde{\nu}'}(\cdot|s_t)$ equals $\widehat{\pi}_t(\cdot|s_t, \nu_{t:T,X})$ for every $s_t \in S$, $\pi_t^{\tilde{\nu}'}(\cdot|s_t)$ is optimal for $\widehat{V}_t(s_t, \nu_{t:T,S})$ for every $s_t \in S$. From this, Lemma 6.3 (i) ensures the claim to hold.

Next we claim that Γ is convex-valued. Let $\nu_{0:T} \in \Xi$, $\nu'_{0:T}, \nu''_{0:T} \in \Gamma(\nu_{0:T})$, and $\lambda \in (0,1)$. For every $t = 0, \dots, T-1$, define $\tilde{\nu}_t \in \mathcal{P}(S \times A)$ by

$$\tilde{\nu}_t(ds_t, da_t) := \lambda \nu'_t(ds_t, da_t) + (1-\lambda)\nu''_t(ds_t, da_t).$$

We claim that $\tilde{\nu}_{0:T} \in \Gamma(\nu_{0:T})$. Since it is straightforward to see that $\tilde{\nu}_{0:T} \in \mathcal{B}(\nu_{0:T})$, we will show that $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$.

It is clear that $\tilde{\nu}_{0,S} = \lambda\nu'_{0,S} + (1-\lambda)\nu''_{0,S} = \mu^o$ (since $\nu'_{0,S} = \nu''_{0,S} = \mu^o$; see Definition 3.8 (i)). Denote by $p_{0:T-1}^{\nu'}$ and $p_{0:T-1}^{\nu''}$ the sequences of kernels enabling $\nu'_{0:T} \in \mathcal{C}(\nu_{0:T})$ and $\nu''_{0:T} \in \mathcal{C}(\nu_{0:T})$ respectively.

Then for every $t = 0, \dots, T-2$, we define $p_t^{\tilde{\nu}} : S \times A \times (\mathcal{P}(S))^{T-t} \to \mathcal{P}(S)$ by for every $(s_t, a_t, \mu_{t:T}) \in S \times A \times (\mathcal{P}(S))^{T-t}$,

$$p_t^{\tilde{\nu}}(\cdot|s_t, a_t, \mu_{t:T}) := \lambda p_t^{\nu'}(\cdot|s_t, a_t, \mu_{t:T}) + (1-\lambda)p_t^{\nu''}(\cdot|s_t, a_t, \mu_{t:T}).$$

Note that for every $t = 0, \dots, T-2$, $p_t^{\nu'}(\cdot|s_t, a_t, \nu_{t:T,S})$, $p_t^{\nu''}(\cdot|s_t, a_t, \nu_{t:T,S}) \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu_{t:T,S})$ for every $(s_t, a_t) \in S \times A$ and $\widehat{\mathfrak{P}}_t$ is convex-valued (see Remark 3.3). Therefore, for every $t = 0, \dots, T-2$, it holds that $p_t^{\tilde{\nu}}(\cdot|s_t, a_t, \nu_{t:T,S}) \in \widetilde{\mathfrak{P}}_t(s_t, a_t, \nu_{t:T,S})$ for every $(s_t, a_t) \in S \times A$.

Furthermore, it also holds for every $t = 0, \dots, T-2$ that

$$\tilde{\nu}_{t+1,S}(\cdot) = \lambda\nu'_{t+1,S}(\cdot) + (1-\lambda)\nu''_{t+1,S}(\cdot) = \int_{S\times A} p_t^{\tilde{\nu}}(\cdot|s_t, a_t, \nu_{t:T,S})\nu_t(ds_t, da_t).$$

We hence have that $\tilde{\nu}_{0:T} \in \mathcal{C}(\nu_{0:T})$. This completes the proof. □

Proof of Proposition 3.9 (ii). Let $(\nu^n_{0:T}, \xi^n_{0:T})_{n\in\mathbb{N}} \subseteq \Xi \times \Xi$ be a sequence such that for every $n \in \mathbb{N}$, $\xi^n_{0:T} \in \Gamma(\nu^n_{0:T})$ and that for every $t = 0, \dots, T-1$ as $n \to \infty$,

$$\nu^n_t \rightharpoonup \nu^\star_t, \quad \xi^n_t \rightharpoonup \xi^\star_t, \tag{6.6}$$

with some $(\nu^\star_{0:T}, \xi^\star_{0:T}) \in \Xi \times \Xi$.

To prove $\mathrm{Gr}(\Gamma)$ is closed, it is sufficient to prove that $\xi^\star_{0:T} \in \Gamma(\nu^\star_{0:T})$.

Step 1. We show that $\xi^\star_{0:T} \in \mathcal{C}(\nu^\star_{0:T})$. Since $\xi^n_{0,S} = \mu^o$ for every $n \in \mathbb{N}$ (due to $\xi^n_{0:T} \in \mathcal{C}(\nu^n_{0:T})$), by (6.6) it holds that $\xi^\star_{0,S} = \mu^o$.

For every $n \in \mathbb{N}$, let $p_{0:T-1}^{\xi^n}$ be a sequence of kernels enabling $\xi^n_{0:T} \in \mathcal{C}(\nu^n_{0:T})$ (see Definition 3.8 (i)). For notational simplicity, set $p^n_{0:T-1} := p^{\xi^n}_{0:T-1}$.

Then for every $n \in \mathbb{N}$ and $t = 0, \dots, T-2$, it holds that

$$\xi^n_{t+1,S}(\cdot) = \int_{S\times A} p^n_t(\cdot|s_t, a_t, \nu^n_{t:T,S})\nu^n_t(ds_t, da_t), \tag{6.7}$$

(due to $\xi^n_{0:T} \in \mathcal{C}(\nu^n_{0:T})$) and that for every $(s_t, a_t) \in S \times A$,

$$\mathbb{P}^n_{t,s_t,a_t} := p^n_t(\cdot|s_t, a_t, \nu^n_{t:T,S}) \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu^n_{t:T,S}). \tag{6.8}$$

Fix any $t \in \{0, \ldots, T-2\}$. Let $(s_t, a_t) \in S \times A$. Since $\mathbb{P}^n_{t,s_t,a_t} \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu^n_{t:T,S})$ for every $n \in \mathbb{N}$ and for every $u = t, \ldots, T-1$, $\nu^n_{u,S} \rightharpoonup \nu^\star_{u,S}$ as $n \to \infty$ (see (6.6)), the compact-valueness and upper-hemicontinuity of the correspondence $\widehat{\mathfrak{P}}_t$ (see Remark 3.3) ensure that there exist a subsequence $(\mathbb{P}^{n_k}_{t,s_t,a_t})_{k\in\mathbb{N}}$ and some $\mathbb{P}_{t,s_t,a_t} \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu^\star_{t:T,S})$ such that

$$\mathbb{P}^{n_k}_{t,s_t,a_t} \rightharpoonup \mathbb{P}_{t,s_t,a_t} \quad \text{as } k \to \infty \tag{6.9}$$

(see [2, Theorem 17.20]). Since both S and A are finite (see Assumption 3.1 (i)), by using the same arguments presented for (6.9) a finite number of times, we can and do choose a subsequence $(\mathbb{P}^{n_k}_{t,s_t,a_t})_{k\in\mathbb{N}}$ of the one in (6.8) and have $(\mathbb{P}_{t,s_t,a_t})_{(s_t,a_t)\in S\times A}$ (for notational simplicity, we do not relabel that sequence) for which (6.9) holds with $\mathbb{P}_{t,x,a} \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu^\star_{t:T,S})$ for every $(s_t, a_t) \in S\times A$.

From this, we can define a mapping

$$p^\star_t : S \times A \ni (s_t, a_t) \mapsto p^\star_t(\cdot|s_t, a_t) := \mathbb{P}_{t,s_t,a_t} \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu^\star_{t:T,S}). \tag{6.10}$$

Lemma 6.1 enables to extend $p^\star_t$ as a stochastic kernel $\overline{p}^\star_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \overline{p}^\star_t(\cdot|s_t, a_t, \mu_{t:T}) \in \widehat{\mathfrak{P}}_t(s_t, a_t, \mu_{t:T})$ such that for every $(s_t, a_t) \in S \times A$, it holds

$$\overline{p}^\star_t(\cdot|s_t, a_t, \nu^\star_{t:T,S}) = p^\star_t(\cdot|s_t, a_t). \tag{6.11}$$

By the consecutive constructions given in (6.10) and (6.11), the limit (6.9) together with (6.8) ensures that for every $(s_t, a_t) \in S \times A$, as $k \to \infty$,

$$p^{n_k}_t(\cdot|s_t, a_t, \nu^{n_k}_{t:T,S}) \rightharpoonup \overline{p}^\star_t(\cdot|s_t, a_t, \nu^\star_{t:T,S}). \tag{6.12}$$

Now we claim that as $k \to \infty$,

$$\int_{S\times A} p^{n_k}_t(\cdot|s_t, a_t, \nu^{n_k}_{t:T,S})\nu^{n_k}_t(ds_t, da_t) \rightharpoonup \int_{S\times A} \overline{p}^\star_t(\cdot|s_t, a_t, \nu^\star_{t:T,S})\nu^\star_t(ds_t, da_t). \tag{6.13}$$

To that end, for every $k \in \mathbb{N}$ denote by $w^{n_k}_t(s_t, a_t)$ and $w^\star_t(s_t, a_t)$ the weights of $\nu^{n_k}_t$ and $\nu^\star_t$ at $(s_t, a_t) \in S \times A$ and by $w^{n_k}_{t,s_t,a_t}(s_{t+1})$ and $\overline{w}^\star_{t,s_t,a_t}(s_{t+1})$ the weights of $p^{n_k}_t(ds_{t+1}|s_t, a_t, \nu^{n_k}_{t:T,S})$ and $\overline{p}^\star_t(ds_{t+1}|s_t, a_t, \nu_{t:T,S})$ at $s_{t+1} \in S$. Then by (6.6) and (6.12) (since S and A are finite; see Assumption 3.1 (i)), it holds that for every $s_t, s_{t+1} \in S$ and $a_t \in A$, as $k \to \infty$,

$$w^{n_k}_t(s_t, a_t) \to w^\star_t(s_t, a_t), \quad w^{n_k}_{t,s_t,a_t}(s_{t+1}) \to \overline{w}^\star_{t,s_t,a_t}(s_{t+1}). \tag{6.14}$$

Let $g : S \to \mathbb{R}$ be any mapping (which is obviously in $C_b(S;\mathbb{R})$ as S is finite). Then since for every $k \in \mathbb{N}$

$$\begin{aligned}&\int_{S\times A}\int_S g(s_{t+1})p^{n_k}_t(ds_{t+1}|s_t, a_t, \nu^{n_k}_{t:T,S})\nu^{n_k}_t(ds_t, da_t)\\ &\quad= \sum_{(s_t,a_t)\in S\times A} w^{n_k}_t(s_t, a_t) \sum_{s_{t+1}\in S} w^{n_k}_{t,s_t,a_t}(s_{t+1})g(s_{t+1}),\end{aligned}$$

from (6.14) (together with the finiteness of S and A), it follows that

$$\begin{aligned}&\lim_{k\to\infty}\int_{S\times A}\int_S g(s_{t+1})p^{n_k}_t(ds_{t+1}|s_t, a_t, \nu^{n_k}_{t:T,S})\nu^{n_k}_t(ds_t, da_t)\\ &\quad= \sum_{(s_t,a_t)\in S\times A} w^\star_t(s_t, a_t) \sum_{s_{t+1}\in S} \overline{w}^\star_{t,s_t,a_t}(s_{t+1})g(s_{t+1})\\ &\quad= \int_{S\times A}\int_S g(s_{t+1})\overline{p}^\star_t(ds_{t+1}|s_t, a_t, \nu^\star_{t:T,S})\nu^\star_t(ds_t, da_t),\end{aligned}$$

which ensures the claim given in (6.13) to hold.

Using (6.13) together with (6.7) and (6.6), we hence have that

$$\xi^\star_{t+1,S}(\cdot) = \int_{S\times A} \overline{p}^\star_t(\cdot|s_t, a_t, \nu^\star_{t:T,S})\nu^\star_t(ds_t, da_t),$$

where we recall that $\overline{p}^\star_t$ satisfies (6.11) for every $(s_t, a_t) \in S \times A$. Since this holds for any $t = 0, \dots, T-2$, we hence have that $\xi^\star_{0:T} \in \mathcal{C}(\nu^\star_{0:T})$.

Step 2. It remains to show that $\xi^\star_{0:T} \in \mathcal{B}(\nu^\star_{0:T})$. Here we follow the arguments of the proof for [51, Proposition 3.9.]. For every $t = 0, \dots, T-1$ and $n \in \mathbb{N}$, set $D^\star_t := D_t(\nu^\star_{t:T})$ and $D^n_t := D_t(\nu^n_{t:T})$ so that $\xi^n_t(D^n_t) = 1$ (because $\xi^n_{0:T} \in \mathcal{B}(\nu^n_{0:T})$; see Definition 3.8 (i)).

Fix any $t = 0, \dots, T-1$. Let $(s^n_t)_{n\in\mathbb{N}} \subseteq S$ and $s_t \in S$ be such that $s^n_t \to s_t$ as $n \to \infty$. Since $\widehat{J}_t(\cdot, \cdot, \nu^n_{t:T,S})\colon S \times A \to \mathbb{R}$ converges continuously[8] to $\widehat{J}_t(\cdot, \cdot, \nu^\star_{t:T,S})\colon S \times A \to \mathbb{R}$ (by Lemma 5.1 (i) and (6.6)) and the action space A is finite, it holds that

$$\lim_{n\to\infty} \max_{a_t\in A} \widehat{J}_t(s^n_t, a_t, \nu^n_{t:T,S}) = \max_{a_t\in A} \widehat{J}_t(s_t, a_t, \nu^\star_{t:T,S}), \tag{6.15}$$

which implies that $\max_{a_t\in A} \widehat{J}_t(\cdot, a_t, \nu^n_{t:T,S})$ converges continuously to $\max_{a_t\in A} \widehat{J}_t(\cdot, a_t, \nu^\star_{t:T,S})$.

For every $M \in \mathbb{N}$, set

$$E^M_t := \left\{(s_t, a_t) \in S \times A \,\middle|\, \max_{a'_t\in A} \widehat{J}_t(s_t, a'_t, \nu^\star_{t:T,S}) \ge \widehat{J}_t(s_t, a_t, \nu^\star_{t:T,X}) + \varepsilon_M\right\} \tag{6.16}$$

to be a closed subset where $(\varepsilon_M)_{M\in\mathbb{N}} \subseteq (0, \infty)$ is a decreasing sequence so that $\lim_{M\to\infty} \varepsilon_M = 0$.

Then since $(D^\star_t)^c = \bigcup_{M=1}^\infty E^M_t$ and $E^M_t \subset E^{M+1}_t$ for every $M \in \mathbb{N}$, the monotone convergence theorem implies that for every $n \in \mathbb{N}$,

$$\begin{aligned} 1 - \xi^n_t(D^\star_t \cap D^n_t) &= \xi^n_t(D^n_t) - \xi^n_t(D^\star_t \cap D^n_t) \\ &= \xi^n_t((D^\star_t)^c \cap D^n_t) = \liminf_{M\to\infty} \xi^n_t(E^M_t \cap D^n_t). \end{aligned}$$

This ensures that

$$\begin{aligned} 1 &= \limsup_{n\to\infty} \liminf_{M\to\infty} \left\{\xi^n_t(D^\star_t \cap D^n_t) + \xi^n_t(E^M_t \cap D^n_t)\right\} \\ &\le \liminf_{M\to\infty} \limsup_{n\to\infty} \left\{\xi^n_t(D^\star_t \cap D^n_t) + \xi^n_t(E^M_t \cap D^n_t)\right\}. \end{aligned} \tag{6.17}$$

We claim that for every $M \in \mathbb{N}$,

$$\limsup_{n\to\infty} \xi^n_t(E^M_t \cap D^n_t) = \limsup_{n\to\infty} \int_{S\times A} \mathbf{1}_{\{(s_t,a_t)\in E^M_t\cap D^n_t\}} \xi^n_t(ds_t, da_t) = 0. \tag{6.18}$$

Fix any $M \in \mathbb{N}$. We firstly show that $\mathbf{1}_{\{(s_t,a_t)\in E^M_t\cap D^n_t\}} : S \times A \mapsto \mathbb{R}$ converges continuously to 0 as $n \to \infty$. Let $(s^n_t, a^n_t)_{n\in\mathbb{N}}$ be a sequence such that $(s^n_t, a^n_t) \to (s^\star_t, a^\star_t) \in E^M_t$ as $n \to \infty$. Then by (6.15) and (6.16),

$$\begin{aligned} \lim_{n\to\infty} \max_{a_t\in A} \widehat{J}_t(x^n_t, a_t, \nu^n_{t:T,S}) &= \max_{a_t\in A} \widehat{J}_t(s^\star_t, a_t, \nu^\star_{t:T,S}) \\ &\ge \widehat{J}_t(s^\star_t, a^\star_t, \nu^\star_{t:T,S}) + \varepsilon_M \\ &= \lim_{n\to\infty} \widehat{J}_t(s^n_t, a^n_t, \nu^n_{t:T,S}) + \varepsilon_M. \end{aligned}$$

Hence, for sufficiently large n, we have $\max_{a_t\in A} \widehat{J}_t(s^n_t, a_t, \nu^n_{t:T,S}) > \widehat{J}_t(s^n_t, a^n_t, \nu^n_{t:T,S})$ which implies that $(s^n_t, a^n_t) \notin D^n_t$. Hence we have that $\mathbf{1}_{\{(s_t,a_t)\in E^M_t\cap D^n_t\}}$ converges continuously to 0 as $n \to \infty$.

[8] Suppose g and $(g_n)_{n\in\mathbb{N}}$ are measurable functions on a metric space E. The sequence $(g_n)_{n\in\mathbb{N}}$ is said to converge to g continuously if $\lim_{n\to\infty} g_n(e_n) = g(e)$ for any sequence $(e_n)_{n\in\mathbb{N}}$ with $e_n \to e \in E$.

From this and the limit $\xi_t^n \rightharpoonup \xi_t^\star$ as $n \to \infty$ (see (6.6)), an application of [53, Theorem 3.3] ensures the claim given in (6.18) to hold for every $M \in \mathbb{N}$.

Combining this with (6.17), we have

$$1 \leq \limsup_{n\to\infty} \xi_t^n(D_t^\star \cap D_t^n) \leq \limsup_{n\to\infty} \xi_t^n(D_t^\star).$$

Furthermore, since $D_t^\star$ is closed, the Portmanteau theorem (see e.g., [7, Theorem 2.1], [9, Theorem 8.2.3]) implies that $\limsup_{n\to\infty} \xi_t^n(D_t^\star) \leq \xi_t^\star(D_t^\star)$. Hence, we have shown that $\xi_t^\star(D_t^\star) = 1$.

Since this holds for any $t = 0, \dots, T-1$, we hence have that $\xi_{0:T}^\star \in \mathcal{C}(\nu_{0:T}^\star)$. This completes the proof. □

Proof of Proposition 3.9 (iii). Note that Ξ is a compact convex topological space. Furthermore, Γ is non-empty, convex-valued and its graph is closed (see Proposition 3.9 (i), (ii)). Therefore, by Kakutani's fixed point theorem (see, e.g., [2, Corollary 17.55, p. 583]), Γ has a fixed point $\nu_{0:T}^*$, i.e., $\nu_{0:T}^* \in \Gamma(\nu_{0:T}^*)$. □

6.3. **Proof of Theorem 3.10.**

Proof of Theorem 3.10. By Proposition 3.9 (iii), Γ has a fixed point $\nu_{0:T}^*$, i.e., $\nu_{0:T}^* \in \Gamma(\nu_{0:T}^*)$.

Then, since $\nu_{0:T}^* \in \mathcal{C}(\nu_{0:T}^*)$, it holds that $\nu_{0,S}^* = \mu^o$. Furthermore, Lemma 6.2 ensures that for every $t = 0, \dots, T-1$, there exists $\overline{p}_t^{\nu^*} : S \times A \times \mathcal{P}(S) \ni (s_t, a_t, \mu_t) \mapsto \overline{p}_t^{\nu^*}(\cdot|s_t, a_t, \mu_t) \in \mathcal{P}(S)$ defined by

$$\overline{p}_t^{\nu^*}(\cdot|s_t, a_t, \mu_t) := \begin{cases} p_t^{\nu^*}(\cdot|s_t, a_t, \mu_t, \nu_{t+1:T,S}^*) & \text{if } \ t \leq T-2; \\ \widehat{p}_t(\cdot|s_t, a_t, \mu_t) & \text{if } \ t = T-1, \end{cases} \tag{6.19}$$

where $\widehat{p}_{0:T}$ is the sequence of the measurable selectors given in Lemma 3.2 (i) and $p_{0:T-1}^{\nu^*}$ is the sequence of the corresponding kernels enabling $\nu_{0:T}^* \in \mathcal{C}(\nu_{0:T}^*)$, i.e., for $t = 0, \dots, T-2$,

$$\begin{aligned} & p_t^{\nu^*}(\cdot|s_t, a_t, \nu_{t:T,S}^*) \in \widehat{\mathfrak{P}}_t(s_t, a_t, \nu_{t:T,S}^*) \ \text{ for every } (s_t, a_t) \in S \times A, \\ & \text{and } \ \nu_{t+1,S}^*(\cdot) = \int_{S\times A} p_t^*(\cdot|s_t, a_t, \nu_{t:T,S}^*)\nu_t^*(ds_t, da_t), \end{aligned} \tag{6.20}$$

(see Definition 3.8), and that for every $t = 0, \dots, T-1$, $\overline{p}_t^{\nu^*}(\cdot|s_t, a_t, \nu_{t,S}^*)$ is optimal for $\widehat{J}_t(s_t, a_t, \nu_{t:T,S}^*)$ for every $(s_t, a_t) \in S \times A$.

Furthermore, since $\nu_{0:T}^* \in \mathcal{B}(\nu_{0:T}^*)$, Lemma 6.3 (ii) ensures that for every $t = 0, \dots, T-1$, there exists $\overline{\pi}_t^{\nu^*} : S \ni s_t \mapsto \overline{\pi}_t^{\nu^*}(\cdot|s_t) \in \mathcal{P}(A)$ defined by

$$\overline{\pi}_t^{\nu^*}(\cdot|s_t) := \begin{cases} \pi_t^{\nu^*}(\cdot|s_t) & \text{if } \ w_t^*(s_t) > 0; \\ \widehat{\pi}_t(\cdot|s_t, \nu_{t:T,S}^*) & \text{else}, \end{cases} \tag{6.21}$$

where $w_t^*(s_t)$ is the weight of $\nu_{t,S}^*$ at $s_t \in S$ and $\widehat{\pi}_{0:T}$ is the sequence of measurable selectors given in Lemma 3.2 (ii), and that for every $t = 0, \dots, T-1$, it holds

$$\nu_t^*(ds_t, da_t) = \overline{\pi}_t^{\nu^*}(da_t|s_t)\nu_{t,S}^*(ds_t), \tag{6.22}$$

and that $\overline{\pi}_t^{\nu^*}(\cdot|s_t)$ is optimal for $\widehat{V}_t(s_t, \nu_{t:T,S}^*)$ for every $s_t \in S$.

The optimality of $\overline{p}_{0:T}^{\nu^*}$ and $\overline{\pi}_{0:T}^{\nu^*}$ ensures that $(\overline{\pi}_{0:T}^{\nu^*}, \overline{p}_{0:T}^{\nu^*})$ is optimal for $V(\nu_{0:T,S}^*)$, i.e., the condition (i) given in Definition 2.3 holds. Furthermore, combining (6.22) with (6.19) and (6.20) ensures that $(\nu_{0:T,S}^*, \overline{\pi}_{0:T}^{\nu^*}, \overline{p}_{0:T}^{\nu^*})$ satisfies condition (ii) given in Definition 2.3 holds. Hence $(\nu_{0:T,S}^*, \overline{\pi}_{0:T}^{\nu^*}, \overline{p}_{0:T}^{\nu^*})$ is a mean-field equilibrium of $(S, A, \mu^o, \mathfrak{P}_{0:T}, r)$. □

7. Proof of results in Section 3.3

7.1. **Proof of Propositions 3.16 and 3.17.**

Let us provide a simple observation that plays an instrumental role in the proof of Proposition 3.16. The proof can be found in Appendix A.

Lemma 7.1. *Let X be a finite space and Y be an arbitrary Borel space. Furthermore, let $(\Lambda_X^{(N)})_{N\in\mathbb{N}}$, $(\widetilde{\Lambda}_X^{(N)})_{N\in\mathbb{N}} \subseteq \mathcal{P}(X)$ be such that for any mapping $f : X \to \mathbb{R}$*

$$\lim_{N\to\infty} \left| \int_X f(x)\Lambda_X^{(N)}(dx) - \int_X f(x)\widetilde{\Lambda}_X^{(N)}(dx) \right| = 0, \tag{7.1}$$

and let $(\Lambda_{Y|X}^{(N)})_{N\in\mathbb{N}}$ be a sequence of stochastic kernels on Y given X such that for every $x \in X$

$$\Lambda_{Y|X}^{(N)}(\cdot|x) \rightharpoonup \Lambda_{Y|X}(\cdot|x) \in \mathcal{P}(Y) \quad \text{as } N \to \infty,$$

where $\Lambda_{Y|X} : X \mapsto \mathcal{P}(Y)$ is another stochastic kernel on Y given X. For $N \in \mathbb{N}$, denote by

$$\begin{aligned}
\Lambda^{(N)}(dx, dy) &:= \Lambda_{Y|X}^{(N)}(dy|x)\Lambda_X^{(N)}(dx) \in \mathcal{P}(X \times Y),\\
\widetilde{\Lambda}_1^{(N)}(dx, dy) &:= \Lambda_{Y|X}^{(N)}(dy|x)\widetilde{\Lambda}_X^{(N)}(dx) \in \mathcal{P}(X \times Y),\\
\widetilde{\Lambda}_2^{(N)}(dx, dy) &:= \Lambda_{Y|X}(dy|x)\widetilde{\Lambda}_X^{(N)}(dx) \in \mathcal{P}(X \times Y).
\end{aligned}$$

Then, for both $i = 1$, 2, we have that for every $g \in C_b(X \times Y)$,

$$\lim_{N\to\infty} \left| \int_{X\times Y} g(x, y)\Lambda^{(N)}(dx, dy) - \int_{X\times Y} g(x, y)\widetilde{\Lambda}_i^{(N)}(dx, dy) \right| = 0.$$

Before we proceed to start proving Proposition 3.16, let us briefly comment on explicit characterizations of the laws and stochastic kernels given in Definition 3.12.

Remark 7.2. Let $(\pi_{0:T}^{(N)})_{N\in\mathbb{N}} \subseteq \Pi$ be a sequence of arbitrary Markov policies. For every $N \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, let $\mathbb{P}^{*|(N)} \in \mathcal{Q}(\mu^*_{0:T}, \pi_{0:T}^{(N)})$ and $\overline{\mathbb{P}}_i^{N|(N)} \in \mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T,i}^{N|(N)})$ (depending on $\pi_{0:T}^{(N)}$) be given in Definition 3.11. Then the following hold for every $t = 0, \dots, T-1$:

(i) The laws $\mathbb{M}_t^{*|(N)}$, $\mathbb{M}_{t,i}^{N|(N)} \in \mathcal{P}(S \times A)$ given in Definition 3.12 (i) are characterized by

$$\begin{aligned}
\mathbb{M}_t^{*|(N)}(ds_t, da_t) &:= \pi_t^{(N)}(da_t|s_t)\mathbb{L}_t^{*|(N)}(ds_t),\\
\mathbb{M}_{t,i}^{N|(N)}(ds_t, da_t) &:= \pi_t^{(N)}(da_t|s_t)\mathbb{L}_{t,i}^{N|(N)}(ds_t),
\end{aligned}$$

where $\mathbb{L}_t^{*|(N)}, \mathbb{L}_{t,i}^{N|(N)} \in \mathcal{P}(S)$ denote the law of s_t under $\mathbb{P}^{*|(N)}$ and the law of s_t^i under $\overline{\mathbb{P}}_i^{N|(N)}$, respectively.

(ii) The stochastic kernel $\mathbb{K}_{t,i}^{N|(N)} : S \times A \ni (s_t, a_t) \mapsto \mathbb{K}_{t,i}^{N|(N)}(ds_{t+1}, d\mu_t|s_t, a_t) \in \mathcal{P}(S \times \mathcal{P}(S))$ given in Definition 3.12 (ii) satisfies that for every $(s_t^i, a_t^i) = (s_t, a_t) \in S \times A$,[9]

$$\begin{aligned}
\mathbb{K}_{t,i}^{N|(N)}(ds_{t+1}, d\mu_t|s_t, a_t) := p_{t,i}^{N|(N),i}\big(ds_{t+1}|(\overline{s}_t^{N,-i}, s_t), (\overline{a}_t^{N,-i}, a_t)\big)\, \overline{\pi}_t^{N-1|*}(d\overline{a}_t^{N,-i}|\overline{s}_t^{N,-i})&\\
\delta_{e^N((\overline{s}_t^{N,-i}, s_t))}(d\mu_t)\, \overline{\mathbb{L}}_{t,i}^{N|(N),-i}(d\overline{s}_t^{N,-i}|s_t)&
\end{aligned}$$

where for every $(\overline{s}_t^N, \overline{a}_t^N) \in S^N \times A^N$,

- $p_{t,i}^{N|(N),i}\big(\,\cdot\,|\overline{s}_t^N, \overline{s}_t^N\big) \in \mathcal{P}(S)$ is the i-th marginal of $\overline{p}_{t,i}^{N|(N)}\big(\,\cdot\,|\overline{s}_t^N, \overline{a}_t^N\big) \in \mathcal{P}(S^N)$;
- $\overline{\pi}_t^{N-1|*}$ is the $N-1$ tuple of π_t^* (as $\overline{\pi}_t^{N|*}$ given in Definition 3.11 (iii));

[9] Denote by $\overline{s}_t^{N,-i} := (s_t^1, \dots, s_t^{i-1}, s_t^{i+1}, \dots, s_t^N) \in S^{N-1}$ the whole agents' state configurations except for the agent i's state s_t^i at time t. The same convention applies to $\overline{a}_t^{N,-i} \in A^{N-1}$. Moreover, as in Footnote 3, we apply the convention therein to $(\overline{s}_t^{N,-i}, s) \in S^N$ and $(\overline{a}_t^{N,-i}, a) \in A^N$.

- $\delta_{e^N(\overline{s}_t^N)} \in \mathcal{P}(\mathcal{P}(S))$ is the Dirac measure on $\mathcal{P}(S)$ at $e^N(\overline{s}_t^N) \in \mathcal{P}(S)$;
- $\overline{\mathbb{L}}_{t,i}^{N|(N),-i} : S \ni s_t \mapsto \overline{\mathbb{L}}_{t,i}^{N|(N),-i}(\cdot|s_t) \in \mathcal{P}(S^{N-1})$ is a stochastic kernel on S^{N-1} given S so that $\overline{\mathbb{L}}_{t,i}^{N|(N),-i}(\cdot|s_t)$ is the conditional law of $\overline{s}_t^{N,-i}$ under $\overline{\mathbb{P}}_i^{N|(N)}$ given $s_t^i = s_t \in S$.

Proof of Proposition 3.16. We note that by Remark 3.13, the notation for $\mathbb{L}_{0:T,i}^{N|(N)}$ (given in Remark 7.2) can be simplified as for every $i = 1, \dots, N$, $\mathbb{L}_{0:T}^{N|(N)} := \mathbb{L}_{0:T,i}^{N|(N)}$. Then it holds that for every $t = 0, \dots, T-1$

$$\mathbb{M}_t^{N|(N)}(ds_t, da_t) = \pi_t^{(N)}(da_t|s_t)\mathbb{L}_t^{N|(N)}(ds_t),$$

where $\mathbb{M}_{0:T}^{N|(N)}$ is given in Remark 3.13.

Let $\mathbb{L}_{0:T}^{*|(N)}$ be given in Remark 7.2 (i). Then we claim that if the following holds for some $t \in \{0, \dots, T-2\}$: for every mapping $f : S \to \mathbb{R}$

$$\lim_{N\to\infty} \left| \int_S f(s_t)\mathbb{L}_t^{*|(N)}(ds_t) - \int_S f(s_t)\mathbb{L}_t^{N|(N)}(ds_t) \right| = 0, \tag{7.2}$$

then the following also holds: for every mapping $f : S \to \mathbb{R}$

$$\lim_{N\to\infty} \left| \int_S f(s_{t+1})\mathbb{L}_{t+1}^{*|(N)}(ds_{t+1}) - \int_S f(s_{t+1})\mathbb{L}_{t+1}^{N|(N)}(ds_{t+1}) \right| = 0. \tag{7.3}$$

Since S is finite (see Assumption 3.1) and the convergence in (7.2) holds, we apply Lemma 7.1 (by setting $\mathbb{L}_t^{*|(N)} \curvearrowright \Lambda_X^{(N)}$, $\mathbb{L}_t^{N|(N)} \curvearrowright \widetilde{\Lambda}_X^{(N)}$, and $\pi_t^{(N)} \curvearrowright \Lambda_{Y|X}^{(N)}$ for every $N \in \mathbb{N}$) to have that for every mapping $h : S \times A \to \mathbb{R}$

$$\lim_{N\to\infty} \left| \int_{S\times A} h(s_t, a_t)\mathbb{M}_t^{*|(N)}(ds_t, da_t) - \int_{S\times A} h(s_t, a_t)\mathbb{M}_t^{N|(N)}(ds_t, da_t) \right| = 0. \tag{7.4}$$

Furthermore, since $S \times A$ is finite (see Assumption 3.1) by the weak convergence given in Assumption 3.14, we apply Lemma 7.1 (together with (7.4) and setting $\mathbb{M}_t^{N|(N)} \curvearrowright \Lambda_X^{(N)}$, $\mathbb{M}_t^{*|(N)} \curvearrowright \widetilde{\Lambda}_X^{(N)}$, $\mathbb{K}_t^{N|(N)} \curvearrowright \Lambda_{Y|X}^{(N)}$, and $p_t^*(ds_{t+1}|\cdot,\cdot,\mu_t)\delta_{\mu_t^*}(d\mu_t) \curvearrowright \Lambda_{Y|X}$ for every $N \in \mathbb{N}$) to have (3.14).

In particular, by Definition 3.12 (iii) and Remark 3.13, the marginals of $\mathbb{Q}_t^{N|(N)}$ and $\mathbb{Q}_t^{*|(N)}$ with respect to s_{t+1} equal $\mathbb{L}_{t+1}^{N|(N)}$ and $\mathbb{L}_{t+1}^{*|N}$, respectively. Hence, (3.14) ensures that (7.3) holds.

Since $\mathbb{L}_0^{N|(N)} = \mathbb{L}_0^{*|(N)} = \mu^o$ for every $N \in \mathbb{N}$, we apply the above claim inductively to have that (3.14) and (7.2) hold for every $t = 0, \dots, T-1$. This completes the proof. □

Proof of Proposition 3.17. For $N \in \mathbb{N}$, let $\mathbb{Q}_{0:T}^{*|(N)}, \mathbb{Q}_{0:T}^{N|(N)}$ be given in Definition 3.12 (iii) and Remark 3.13, respectively. Since the following hold for every $t = 0, \dots, T-1$ that

$$\mathbb{E}^{\overline{\mathbb{P}}_1^{N|(N)}}\left[r(s_t^1, a_t^1, s_{t+1}^1, e_t^N(\overline{s}_t^N))\right] = \int_{S\times A\times S\times\mathcal{P}(S)} r(s_t, a_t, s_{t+1}, \mu_t)\mathbb{Q}_t^{N|(N)}(ds_t, da_t, ds_{t+1}, d\mu_t),$$

$$\mathbb{E}^{\mathbb{P}^{*|(N)}}\left[r(s_t, a_t, s_{t+1}, \mu_t^*)\right] = \int_{S\times A\times S\times\mathcal{P}(S)} r(s_t, a_t, s_{t+1}, \mu_t)\mathbb{Q}_t^{*|(N)}(ds_t, da_t, ds_{t+1}, d\mu_t)$$

with $\mathbb{P}^{*|(N)} \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^{(N)})$ and $\overline{\mathbb{P}}_1^{N|(N)} \in \mathcal{Q}^N(\mu^o, \overline{\pi}_{0:T}^{N|(N)})$ given in Definition 3.11, Proposition 3.16 (together with $r \in C_b(S \times A \times S \times \mathcal{P}(S))$; see Assumption 3.1 (iii)) ensures that for every $t = 0, \dots, T-1$

$$\lim_{N\to\infty} \left| \mathbb{E}^{\overline{\mathbb{P}}_1^{N|(N)}}\left[r(s_t^1, a_t^1, s_{t+1}^1, e_t^N(\overline{s}_t^N))\right] - \mathbb{E}^{\mathbb{P}^{*|(N)}}\left[r(s_t, a_t, s_{t+1}, \mu_t^*)\right] \right| = 0.$$

Hence,

$$\lim_{N\to\infty} \left| J_1^N(\mu^o, \overline{\pi}_{0:T,1}^{N|(N)}) - \mathbb{E}^{\mathbb{P}^{*|(N)}}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t^*)\right] \right|$$

$$
\begin{aligned}
&= \lim_{N\to\infty} \left| \sum_{t=0}^{T-1} \mathbb{E}^{\overline{\mathbb{P}}_1^{N|(N)}}\big[r(s_t^1, a_t^1, s_{t+1}^1, e_t^N(\overline{s}_t^N))\big] - \sum_{t=0}^{T-1} \mathbb{E}^{\mathbb{P}^{*|(N)}}\big[r(s_t, a_t, s_{t+1}, \mu_t^*)\big] \right| \\
&\leq \sum_{t=0}^{T-1} \lim_{N\to\infty} \left| \mathbb{E}^{\overline{\mathbb{P}}_1^{N|(N)}}\big[r(s_t^1, a_t^1, s_{t+1}^1, e_t^N(\overline{s}_t^N))\big] - \mathbb{E}^{\mathbb{P}^{*|(N)}}\big[r(s_t, a_t, s_{t+1}, \mu_t^*)\big] \right| = 0.
\end{aligned}
$$

This completes the proof. □

7.2. Proof of Theorem 3.19.

Proof of Theorem 3.19. Let $\varepsilon > 0$. By using the same arguments presented in Remark 3.13, it is enough to show that there exists $N(\varepsilon) \in \mathbb{N}$ such that for each $N \geq N(\varepsilon)$,

$$J_1^N(\mu^o, \overline{\pi}_{0:T}^{N|*}) + \varepsilon \geq \sup_{\pi_{0:T}\in\Pi} J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T})),$$

where J_1^N denotes the worst-case reward for agent $i = 1$.

For each $N \geq \mathbb{N}$, let $\pi_{0:T}^{(N)} \in \Pi$ be a sequence of policies satisfying that

$$J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T}^{(N)})) > \sup_{\pi_{0:T}\in\Pi} J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T})) - \frac{\varepsilon}{3}. \tag{7.5}$$

By Proposition 3.4 (ii) (by replacing $\tilde{\mu}_{0:T}$ as $\mu_{0:T}^*$; see (3.11)) it holds that

$$\sup_{\pi_{0:T}\in\Pi} \mathbb{E}^{\mathbb{P}(\mu_{0:T}^*, \pi_{0:T}, p_{0:T}^*)}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t^*)\right] = \mathbb{E}^{\mathbb{P}^*}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t^*)\right] = V(\mu_{0:T}^*),$$

where $\mathbb{P}^* = \mathbb{P}(\mu_{0:T}^*, \pi_{0:T}^*, p_{0:T}^*) \in \mathcal{Q}(\mu_{0:T}^*, \pi_{0:T}^*)$ (see Definition 3.11 (i)).

Moreover since $J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T}^{(N)})) = J_1^N(\mu^o, \overline{\pi}_{0:T,1}^{N|(N)})$ and $\mathbb{P}(\mu_{0:T}^*, \pi_{0:T}^{(N)}, p_{0:T}^*) = \mathbb{P}^{*|(N)}$ (see Definition 3.11), we apply Proposition 3.17 to have

$$
\begin{aligned}
\lim_{N\to\infty} J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T}^{(N)})) &= \lim_{N\to\infty} \mathbb{E}^{\mathbb{P}(\mu_{0:T}^*, \pi_{0:T}^{(N)}, p_{0:T}^*)}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t^*)\right] \\
&\leq \sup_{\pi_{0:T}\in\Pi} \mathbb{E}^{\mathbb{P}(\mu_{0:T}^*, \pi_{0:T}, p_{0:T}^*)}\left[\sum_{t=0}^{T-1} r(s_t, a_t, s_{t+1}, \mu_t^*)\right] \\
&= V(\mu_{0:T}^*).
\end{aligned} \tag{7.6}
$$

Combining (7.5)–(7.6) and Remark 3.18, we can choose $N(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N(\varepsilon)$

$$
\begin{aligned}
\sup_{\pi_{0:T}\in\Pi} J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T})) - \varepsilon &< J_1^N(\mu^o, (\overline{\pi}_{0:T}^{N|*,-1}, \pi_{0:T}^{(N)})) - \frac{2\varepsilon}{3} \\
&\leq V(\mu_{0:T}^*) - \frac{\varepsilon}{3} \\
&\leq J_1^N(\mu^o, \overline{\pi}_{0:T}^{N|*}).
\end{aligned}
$$

This completes the proof. □

Appendix A. Supplementary proofs

Proof of Lemma 4.2. Fix arbitrary $\lambda \geq 0$ and $c > 0$. We first claim that $\mathfrak{P}_{0:T}^{\lambda}$ satisfies Assumption 3.1 (ii). Let $t \in \{0, \dots, T-1\}$, and let $s_t \in S$, $a_t \in A$, and $\mu_t, \tilde{\mu}_t \in \mathcal{P}(S)$ be arbitrarily chosen. Since the reference kernel p^o does not depend on the argument μ and hence, $\mathfrak{P}_t^{\lambda}(s_t, a_t, \mu_t) = \mathfrak{P}_t^{\lambda}(s_t, a_t, \tilde{\mu}_t)$. Furthermore, as $\mathfrak{P}_t^{\lambda}(s_t, a_t, \mu_t)$ is a 1-Wasserstein ball around $p^o(\cdot \mid s_t, a_t, \mu_t)$, it is clearly non-empty, convex-valued, compact-valued.

Furthermore, since $\mathfrak{P}_t^\lambda(s_t, a_t, \mu_t) = \mathfrak{P}_t^\lambda(s_t, a_t, \widetilde{\mu}_t)$, for any $\mathbb{P} \in \mathfrak{P}_t^\lambda(s_t, a_t, \mu_t)$, we choose the same one $\widetilde{\mathbb{P}} := \mathbb{P} \in \mathfrak{P}_t^\lambda(s_t, a_t, \widetilde{\mu}_t)$ to get

$$0 = d_{W_1}(\mathbb{P}, \widetilde{\mathbb{P}}) \leq d_{W_1}(\mu, \widetilde{\mu}).$$

It remains to show that $\mathfrak{P}_t^\lambda$ is continuous (i.e., upper- and lower-hemicontinuous). To that end, consider an arbitrary sequence[10]

$$\big((s_n, a_n, \mu_n), \mathbb{P}_n\big)_{n\in\mathbb{N}} \subseteq \mathrm{Gr}(\mathfrak{P}_t^\lambda)$$

such that $(s_n, a_n) \to (s, a)$ and $\mu_n \rightharpoonup \mu$ as $n \to \infty$.

Since S and A are finite, there exists $N \in \mathbb{N}$ such that for every $n \geq N$ it holds that $(s_n, a_n, \mu_n) = (s, a, \mu_n)$. Hence, $\mathbb{P}_n \in \mathfrak{P}_t^\lambda(s, a, \mu_n) = \mathfrak{P}_t^\lambda(s, a, \mu)$ for every $n \geq N$. Moreover, since $\mathfrak{P}_t^\lambda(s, a, \mu)$ is compact, there exists a subsequence $(\mathbb{P}_{n_k})_{k\in\mathbb{N}} \subseteq (\mathbb{P}_n)_{n\in\mathbb{N}}$ with $\mathbb{P}_{n_k} \rightharpoonup \mathbb{P} \in \mathfrak{P}_t^\lambda(s, a, \mu)$ as $k \to \infty$. Thus, by [2, Theorem 17.20], $\mathfrak{P}_t^\lambda$ is upper-hemicontinuous.

Again, consider an arbitrary sequence $((s_n, a_n, \mu_n))_{n\in\mathbb{N}} \subset S \times A \times \mathcal{P}(S)$ such that $(s_n, a_n) \to (s, a)$ and $\mu_n \rightharpoonup \mu$ as $n \to \infty$ and let $\mathbb{P} \in \mathfrak{P}_t^\lambda(s, a, \mu)$. As before, there exists $N \in \mathbb{N}$ such that for every $n \geq N$ it holds that $(s_n, a_n, \mu_n) = (s, a, \mu_n)$. Define a sequence $(\mathbb{P}_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}(S)$ by setting

$$\mathbb{P}_n := \begin{cases} p^o(\cdot \mid s_n, a_n, \mu_n) & \text{if } n < N, \\ \mathbb{P} & \text{else.} \end{cases}$$

Then, $\mathbb{P}_n \in \mathfrak{P}_t^\lambda(s_n, a_n, \mu_n)$ for all $n \in \mathbb{N}$ and $\mathbb{P}_n \rightharpoonup \mathbb{P}$ as $n \to \infty$. Hence, by [2, Theorem 17.21], $\mathfrak{P}_t^\lambda$ is lower-hemicontinuous. Hence $\mathfrak{P}_{0:T}^\lambda$ satisfies Assumption 3.1 (ii), as claimed.

We now claim that r given in Definition 4.1 (ii) satisfies Assumption 3.1 (iii).

Since $|\hat{s}| < 4$ and $|a| < 1$ for every $(\hat{s}, a) \in S \times A$ (noting that $S = \{0, 1, \ldots, 4\}$ and $A = \{-1, 0, 1\}$; Definition 4.1), there exists a constant $C_r := \frac{17}{4} + \max\big\{-\log(c), \log(1 + c)\big\} > 0$ satisfying that for every $s, \hat{s} \in S$, $a \in A$, and $\mu \in \mathcal{P}(S)$,

$$\begin{aligned} |r(s, a, \hat{s}, \mu)| &\leq \left|1 - \frac{1}{2}|\hat{s} - 2|\right| + \frac{|a|}{4} + |\log(\mu(\hat{s}) + c)| \\ &\leq 1 + \frac{1}{2}(|\hat{s}| + 2) + \frac{1}{4} + \max\big\{-\log(c), \log(1 + c)\big\} \leq C_r. \end{aligned}$$

Moreover, there exists $L_r := 1/c > 0$ satisfying that for every $s, \hat{s} \in S$, $a \in A$, and $\mu, \hat{\mu} \in \mathcal{P}(S)$,

$$\begin{aligned} |r(s, a, \hat{s}, \mu) - r(s, a, \hat{s}, \hat{\mu})| &= |\log(\hat{\mu}(\hat{s}) + c) - \log(\mu(\hat{s}) + c)| \\ &= \left|\log\left(1 + \frac{\hat{\mu}(\hat{s}) + c}{\mu(\hat{s}) + c} - 1\right)\right| \leq \left|\frac{\hat{\mu}(\hat{s}) + c}{\mu(\hat{s}) + c} - 1\right| \\ &= \frac{1}{\mu(\hat{s}) + c}|\hat{\mu}(\hat{s}) - \mu(\hat{s})| \leq L_r|\hat{\mu}(\hat{s}) - \mu(\hat{s})| \\ &\leq L_r d_{W_1}(\mu, \hat{\mu}). \end{aligned}$$

Hence, r satisfies Assumption 3.1 (iii), as claimed. □

Proof of Lemma 6.1. By the existence of measurable selectors given in Lemma 3.2 (i), we can and do choose a stochastic kernel $p_t'\colon S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto p_t'(\cdot|s_t, a_t, \mu_{t:T}) \in \mathcal{P}(S)$. Then define $\overline{p}_t : S \times A \times (\mathcal{P}(S))^{T-t} \ni (s_t, a_t, \mu_{t:T}) \mapsto \overline{p}_t(\cdot|s_t, a_t, \mu_{t:T}) \in \mathcal{P}(S)$ by

$$\overline{p}_t(\cdot|s_t, a_t, \mu_{t:T}) = \begin{cases} p_t(\cdot|s_t, a_t) & \text{if } \quad \mu_{t:T} = \tilde{\mu}_{t:T}, \\ p_t'(\cdot|s_t, a_t, \mu_{t:T}) & \text{else.} \end{cases} \tag{A.1}$$

[10] We denote by $\mathrm{Gr}(\mathfrak{P}_t^\lambda)$ the graph of $\mathfrak{P}_t^\lambda$.

It is sufficient to show that $\overline{p}_t$ is Borel-measurable. To that end, recall that $\mathcal{B}_{\mathcal{P}(S)}$ and $\mathcal{B}_{S\times A\times(\mathcal{P}(S))^{T-t}}$ denote the Borel σ-field of $\mathcal{P}(S)$ and $S\times A\times(\mathcal{P}(S))^{T-t}$, respectively.

Let $E\in\mathcal{B}_{\mathcal{P}(S)}$. Then since

$$\begin{aligned}\overline{p}_t^{-1}(E)&=\Big\{(s_t,a_t,\mu_{t:T})\in S\times A\times(\mathcal{P}(S))^{T-t}\;\Big|\;\overline{p}_t(\cdot|s_t,a_t,\mu_{t:T})\in E\Big\}\\&=\Big\{(s_t,a_t,\mu_{t:T})\in X\times A\times\{\tilde{\mu}_{t:T}\}\;\Big|\;\overline{p}_t(\cdot|s_t,a_t,\mu_{t:T})\in E\Big\}\\&\qquad\cup\Big\{(s_t,a_t,\mu_{t:T})\in S\times A\times(\mathcal{P}(S))^{T-t}\setminus\{\tilde{\mu}_{t:T}\})\;\Big|\;\overline{p}_t(\cdot|s_t,a_t,\mu_{t:T})\in E\Big\}=:E_1\cup E_2,\end{aligned}$$

we will show that $E_1,\,E_2\in\mathcal{B}_{S\times A\times(\mathcal{P}(S))^{T-t}}$.

Note that by (A.1),

$$\begin{aligned}E_1&=\Big\{(s_t,a_t)\in S\times A\;\Big|\;p_t(\cdot|s_t,a_t)\in E\Big\}\times\{\tilde{\mu}_{t:T}\},\\E_2&=\Big\{(s_t,a_t,\mu_{t:T})\in S\times A\times(\mathcal{P}(S))^{T-t}\;\Big|\;p_t'(\cdot|s_t,a_t,\mu_{t:T})\in E\Big\}\\&\qquad\setminus\Big(\Big\{(s_t,a_t)\in S\times A\;\Big|\;p_t'(\cdot|s_t,a_t,\tilde{\mu}_{t:T})\in E\Big\}\times\{\tilde{\mu}_{t:T}\}\Big)=:E_{2,1}\setminus E_{2,2}.\end{aligned}$$

Since S and A are finite (see Assumption 3.1 (i)), p_t is Borel-measurable. Hence this implies that $E_1\in\mathcal{B}_{S\times A\times(\mathcal{P}(S))^{T-t}}$. For the same reason, it follows that $E_{2,2}\in\mathcal{B}_{S\times A\times(\mathcal{P}(S))^{T-t}}$. Furthermore, since p_t' is Borel-measurable, $E_{2,1}\in\mathcal{B}_{S\times A\times(\mathcal{P}(S))^{T-t}}$. □

Proof of Lemma 7.1. We only prove for $i=2$, as the proof for $i=1$ follows the same line of reasoning. For every $N\in\mathbb{N}$, denote by $w^{(N)}(x)$ the weight representing of the $x\in X$ under $\Lambda_X^{(N)}$, and similarly for $\widetilde{w}^{(N)}(x)$ under $\widetilde{\Lambda}_X^{(N)}$. Let $g\in C_b(X\times Y)$. By the triangle inequality,

$$\left|\int_{X\times Y}g(x,y)\Lambda^{(N)}(dx,dy)-\int_{X\times Y}g(x,y)\widetilde{\Lambda}_2^{(N)}(dx,dy)\right|\le\mathrm{I}^{(N)}+\mathrm{II}^{(N)},$$

where $\mathrm{I}^{(N)}$ and $\mathrm{II}^{(N)}$ are given by

$$\begin{aligned}\mathrm{I}^{(N)}&:=\left|\int_X\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)\Lambda_X^{(N)}(dx)-\int_X\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)\widetilde{\Lambda}_X^{(N)}(dx)\right|,\\\mathrm{II}^{(N)}&:=\left|\int_X\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)\widetilde{\Lambda}_X^{(N)}(dx)-\int_X\int_Y g(x,y)\Lambda_{Y|X}(dy|x)\widetilde{\Lambda}_X^{N}(dx)\right|.\end{aligned}$$

We claim that $\mathrm{I}^{(N)}$ and $\mathrm{II}^{(N)}$ vanish as $N\to\infty$. Indeed, note that for every $N\in\mathbb{N}$

$$\begin{aligned}\mathrm{I}^{(N)}&=\left|\sum_{x\in X}w^{(N)}(x)\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)-\sum_{x\in X}\widetilde{w}^{(N)}(x)\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)\right|\\&\le\sum_{x\in X}\left|w^{(N)}(x)-\widetilde{w}^{(N)}(x)\right|\int_Y|g(x,y)|\Lambda_{Y|X}^{(N)}(dy|x)<C_g\cdot\sum_{x\in X}\left|w^{(N)}(x)-\widetilde{w}^{(N)}(x)\right|,\end{aligned}$$

where $C_g=\sup_{x,y\in X}|g(x,y)|<\infty$ (hence not depending on $N\in\mathbb{N}$) as $g\in C_b(X\times Y)$.

In particular, from the convergence given in (7.1), the finiteness of the space X ensures that $\sum_{x\in X}\left|w^{(N)}(x)-\widetilde{w}^{(N)}(x)\right|\to0$ as $N\to\infty$. Therefore $\mathrm{I}^{(N)}$ vanishes as $N\to\infty$.

And similarly, since $\Lambda_{Y|X}^{(N)}(\cdot|x)\rightharpoonup\Lambda_{Y|X}(\cdot|x)$ as $N\to\infty$ for every $x\in X$ and the space X is finite, we can conclude that

$$\lim_{N\to\infty}\mathrm{II}^{(N)}\le\sum_{x\in X}\widetilde{w}^{(N)}(x)\left(\lim_{n\to\infty}\left|\int_Y g(x,y)\Lambda_{Y|X}^{(N)}(dy|x)-\int_Y g(x,y)\Lambda_{Y|X}(dy|x)\right|\right)=0.$$

This completes the proof. □

References

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

[2] C. D. Aliprantis and K. C. Border. *Infinite dimensional analysis: A Hitchhiker's Guide*. Springer, 2006.

[3] A. Aurell, R. Carmona, G. Dayanikli, and M. Laurière. Optimal incentives to mitigate epidemics: a Stackelberg mean field game approach. *SIAM J. Control Optim.*, 60(2):S294–S322, 2022.

[4] N. Bäuerle. Mean field Markov decision processes. *Appl. Math. Optim.*, 88(1):12, 2023.

[5] D. Bauso, H. Tembine, and T. Başar. Robust mean field games. *Dynam. Games Appl.*, 6(3):277–303, 2016.

[6] A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*, volume 101. New York: Springer-Verlag, 2013.

[7] P. Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

[8] A. Biswas. Mean field games with ergodic cost for discrete time markov processes. *arXiv preprint arXiv:2012.05237*, 2015.

[9] V. I. Bogachev. *Measure Theory: Volume II.* Springer, 2007.

[10] P. Cardaliaguet. *Notes on mean field games (from P.-L. Lions' lectures at Colleġe de France).* Lecture notes, April–May 2010, Tor Vergata, Rome, 2011.

[11] R. Carmona. Applications of mean field games in financial engineering and economic theory. *arXiv preprint arXiv:2012.05237*, 2020.

[12] R. Carmona and F. Delarue. *Probabilistic theory of mean field games with applications I-II.* Springer, 2018.

[13] R. Carmona, F. Delarue, and D. Lacker. Mean field games of timing and models for bank runs. *Appl. Math. Optim.,*, 76:217–260, 2017.

[14] R. Carmona, J.-P. Fouque, and L.-H. Sun. Mean field games and systemic risk. *Commun. Math. Sci.*, 13(4):911–933, 2015.

[15] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *Ann. Appl. Probab.*, 33(6B):5334–5381, 2023.

[16] Z. Chen and L. Epstein. Ambiguity, risk, and asset returns in continuous time. *Econometrica*, 70(4):1403–1443, 2002.

[17] F. Delarue, D. Lacker, and K. Ramanan. From the master equation to mean field game limit theory. *Ann. Probab.*, 48(1):211–263, 2020.

[18] J. Dow and S. R. da Costa Werlang. Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica*, pages 197–204, 1992.

[19] R. Elie, E. Hubert, and G. Turinici. Contact rate epidemic control of COVID-19: an equilibrium view. *Math. Model. Nat. Phenom.*, 15:35, 2020.

[20] R. Elie, J. Pérolat, M. Laurière, M. Geist, and O. Pietquin. On the convergence of model free learning in mean field games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7143–7150, 2020.

[21] R. Elliott, X. Li, and Y.-H. Ni. Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, 49(11):3222–3233, 2013.

[22] L. G. Epstein and T. Wang. Intertemporal asset pricing under Knightian uncertainty. *Econometrica*, pages 283–322, 1994.

[23] N. Gast and B. Gaujal. A mean field approach for optimization in discrete time. *Discrete Event Dyn. Syst.*, 21(1):63–101, 2011.

[24] N. Gast, B. Gaujal, and J.-Y. Le Boudec. Mean field for Markov decision processes: from discrete to continuous optimization. *IEEE. Trans. Autom. Control*, 57(9):2266–2280, 2012.

[25] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *J. Math. Econ.*, 18(2):141–153, 1989.

[26] D. A. Gomes, J. Mohr, and R. R. Souza. Discrete time, finite state space mean field games. *J. Math. Pures Appl.*, 93(3):308–328, 2010.

[27] D. A. Gomes, J. Mohr, and R. R. Souza. Continuous time, finite state mean field games. *Appl. Math. Optim.*, 68(1):99–143, 2013.

[28] D. A. Gomes and J. Saúde. Mean field games models—a brief survey. *Dynam. Games Appl.*, 4:110–154, 2014.

[29] H. Gu, X. Guo, X. Wei, and R. Xu. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM J. Math. Data Sci.*, 3(4):1168–1196, 2021.

[30] H. Gu, X. Guo, X. Wei, and R. Xu. Dynamic programming principles for mean-field controls with learning. *Oper. Res.*, 71(4):1040–1054, 2023.

[31] J. Huang and M. Huang. Mean field LQG games with model uncertainty. In *CDC 2013*. Florence, Dec. 2013.

[32] K. Huang, X. Chen, X. Di, and Q. Du. Dynamic driving and routing games for autonomous vehicles on networks: A mean field game approach. *Transp. Res. C Emerg. Technol.*, 128:103189, 2021.

[33] M. Huang. Large-population LQG games involving a major player: the nash certainty equivalence principle. *SIAM J. Control Optim.*, 48(5):3318–3353, 2010.

[34] M. Huang, R. P. Malhamé, and P. E. Caines. Large population stochastic dynamic games: Closed loop McKean-Vlasov sysyems and the Nash certainity equivalence principle. *Commun. Inf. Syst.*, 6(3):221–252, 2006.

[35] B. Jovanovic and R. W. Rosenthal. Anonymous sequential games. *J. Math. Econ.*, 17(1):77–87, 1988.

[36] A. Lachapelle and M.-T. Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transp. Res. B Methodol.*, 45(10):1572–1589, 2011.

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

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

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

[40] J.-M. Lasry and P.-L. Lions. Mean field games. *Japan. J. Math.*, 2(1):229–260, 2007.

[41] M. Laurière, S. Perrin, J. Pérolat, S. Girgin, P. Muller, R. Elie, M. Geist, and O. Pietquin. Learning in mean field games: A survey. *arXiv*, 2205.12944.

[42] M. Laurière and L. Tangpi. Convergence of large population games to mean field games with interaction through the controls. *SIAM J. Math. Anal.*, 54(3):3535–3574, 2022.

[43] J. Moon and T. 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.

[44] J. Moon and T. Başar. Linear quadratic risk-sensitive and robust mean field games. *IEEE. Trans. Autom. Control*, 62(3):1062–1077, 2016.

[45] J. Moon and T. Başar. Robust mean field games for coupled markov jump linear systems. *Internat. J. Control*, 89(7):1367–1381, 2016.

[46] J. Moon and T. 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, 2016.

[47] M. Motte and H. Pham. Mean-field Markov decision processes with common noise and open-loop controls. *Ann. Appl. Probab.*, 32(2):1421–1458, 2022.

[48] A. Neufeld and J. Sester. Non-concave distributionally robust stochastic control in a discrete time finite horizon setting. *arXiv preprint arXiv:2404.05230*, 2024.

[49] A. Neufeld, J. Sester, and M. Šikić. Markov decision processes under model uncertainty. *Math. Finance*, 33(3):618–665, 2023.

[50] M. Nourian and G. N. Nair. Linear-quadratic-gaussian mean field games under high rate quantization. In *CDC 2013*. Florence, Dec. 2013.

[51] N. Saldi, T. Başar, and M. Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM J. Control Optim.*, 56(6):4256–4287, 2018.

[52] N. Saldi, T. Başar, and M. Raginsky. Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Math. Oper. Res.*, 44(3):1006–1033, 2019.

[53] R. Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhya Ser.A*, pages 380–402, 1982.

[54] H. Tembine, Q. Zhu, and T. Başar. Risk-sensitive mean-field games. *IEEE. Trans. Autom. Control*, 59(4):835–850, 2013.

Insitute of Actuarial and Financial Mathematics & House of Insurance, Leibniz Universität Hannover
Email address: johannes.langner@insurance.uni-hannover.de

Division of Mathematical Sciences, Nanyang Technological University
Email address: ariel.neufeld@ntu.edu.sg

Division of Mathematical Sciences, Nanyang Technological University
Email address: kyunghyun.park@ntu.edu.sg