Dynamic Programming Principles for Learning MFCs

Haotian Gu * Xin Guo * Xiaoli Wei * Renyuan Xu †

November 17, 2019

Abstract

This paper establishes the time consistent property, i.e., the dynamic programming principle (DPP), for learning mean field controls (MFCs). The key idea is to define the correct form of the Q function, called the IQ function, for learning MFCs. This particular form of IQ function reflects the essence of MFCs and is an "integration" of the classical Q function over the state and action distributions. The DPP in the form of the Bellman equation for this IQ function generalizes the classical DPP of Q-learning to the McKean-Vlasov system. It also generalizes the DPP for MFCs controls to the learning framework. In addition, to accommodate model-based learning for MFCs, the DPP for the associated value function is derived. Finally, numerical experiments are presented to illustrate the time consistency of this IQ function.

1 Introduction

MFC/MKV controls. McKean-Vlasov (MKV) processes are stochastic processes governed by stochastic differential equations whose coefficients depend on distributions of the solutions. They were first studied by Henry McKean in 1966 [24]. Nowadays, MKV processes are broadly applied to model collective behaviors of stochastic systems with a large number of mutually interacting agents. MKV controls concern the optimal control of such systems where exchangeable agents interact through the empirical measure of their states. As such, MKV controls are often called mean-field-type controls (MFCs) and are closely related to Mean-Field Games (MFGs). (See [20] and [16] and the references therein).

Mathematically, MFC is the limiting regime of games with infinite number of players, and studies the Pareto optimality for such collaborative games. Therefore analyzing MFCs is simpler than directly studying the large stochastic systems, with good approximations of the latter [19]. In addition to the rapid development on theory of MKV systems and MFCs, [3, 7], there is a growing literature on their applications, including the MKV system in [13] for systemic risk assessment, MKV controls in [25] for a large benevolent planner such as government or the central bank to control taxes or interest rates, and MKV controls in [1] for consumers to choose between new energy resources such as solar panels and traditional ones.

*Department of Industrial Engineering & Operations Research, University of California, Berkeley, USA. **Email:** {haotian_gu, xinguo, xiaoliwei }@berkeley.edu

†Mathematical Institute, University of Oxford, UK. **Email:** xur@maths.ox.ac.uk

MKV/MFC and time inconsistency. Dynamics of most controlled stochastic systems, unless with clear underlying physics principles, are unknown *a priori*. Examples abound: multi-player online role-playing games [17], high frequency tradings [22], and the sharing economy[15], to name a few. Thus it is necessary to consider both learning and control simultaneously. However, literature on learning MFCs virtually does not exist, except for some very recent works of [8], [9] and [29]. They design learning algorithms with the assumption of the dynamics programming principle (DPP) (i.e., the Bellman equation) in their learning frameworks. Indeed, it appears natural to assume DPP for MFCs: MFCs are similar to classical control problems and Markov decision problems (MDPs), both of which rely on the well-established DPP for analysis. However, MFCs differ fundamentally from classical controls or MDPs in that parameters in MKV systems depend on both the marginal distributions of the state and the control. In fact, it has been well recognized that DPP in general does not hold for the controlled MKV system due to its non-Markovian nature ([2], [5], and [6]). That is, MFC problems are inherently time inconsistent. This time inconsistency issue for MFCs has only recently been resolved in a series of papers by considering appropriately enlarged state spaces, including [21] and [26] for a finite time horizon MFCs and [11] for a general MFC framework.

Our Work: time consistency in learning MFCs. Nevertheless, the plague of this time inconsistency issue persists in designing learning algorithms for MFCs. In particular, Example 3.1 in Section 3.1 shows that with a misspecified Q function, MFC problem would be time inconsistent again: with different initial actions the Q-learning algorithm converges to different values. Time consistency property is critical for various RL algorithms. For model-free learning, consistency of Q function is essential for algorithms including Q-learning ([31]) and Actor-Critic method ([18]). For model-based learning, consistency of the value function is the foundation for algorithms involving value iteration and policy evaluation ([12]).

In this paper, we resolve this time inconsistency issue. For model-free learning, our focus is the Q function; for model-based learning, we consider the value function. We will rigorously establish DPP for both an appropriately-defined Q function and its corresponding value functions.

Our first step is to define the correct form of Q function for learning MFCs. To differentiate it from the classical Q function, we call it an IQ function. This particular form of IQ function reflects the essence of MFC (or MKV control problem) as a collaborative game and as the central controller's control problem. A central controller is to coordinate efforts from each individual agent for social optimality. Mathematically, it is an "integration" of the classical Q function over distributions of the state and action from each individual, hence the name "integrated Q (IQ)" function. (See Section 3.4). We then establish the DPP in the form of the Bellman equation for this IQ function. This DPP generalizes the classical DPP of Q-learning for the MDP setting to the MKV system. (See [30] [31], and the standard references [4] and [27] for the classical DPP). It also generalizes the DPP for MFCs controls ([21], [26], and [11]) to the learning framework. In addition, to accommodate model-based learning for MFCs, we establish the DPP for the value function. Finally, we illustrate through numerical experiments the time consistency of this IQ function.

Our contribution. Our work of DPP for learning MFCs differs in two key aspects from [21], [26], and [11] on DPP for value functions of MFCs: one is the aforementioned IQ

function for the learning environment, another is the incorporation of relaxed controls instead of strict control adopted in these works. Relaxed control is a larger set of strategies often encountered in the optimization and learning literature. Relaxed controls are essential for reinforcement learnings which are characterized with exploration and exploitation [27]. Indeed, with limited information of the system, it is necessary to explore and consider a randomized policy with actions/controls sampled from a distribution of actions. (See [32, 10], and [23]). This kind of randomized strategies, also known as mixed strategies for game theory, is exactly the *relaxed control* in control theory. (See [33] and also some recent work [28]). From a control perspective, relaxed controls are also necessary. In classic controls with concave reward functions, the optimal control is necessarily a pure control even for MFGs and MKV controls [19]. However, for large-scale optimization and machine learning problems, l_1 and l_2 norms or entropy terms of the action/state distributions are often introduced either for regularization purpose or to encourage exploration. Consequently, the value function is neither convex nor concave and the optimal control may be a relaxed type.

To the best of our knowledge, this is the first time DPP is rigorously established for learning MFCs, both for the IQ function and for the value function.

Outline of the paper. The rest of the paper is organized as follows. Section 2 formulates the discrete time MFCs problem, with preliminary analysis. Section 3 introduces the IQ function for learning MFCs and establishes corresponding DPP in the form of Bellman functions for both the IQ function and the value function. The paper concludes by revisiting the motivating example 3.1 with the performance of the IQ function.

2 Learning MFC/McKean-Vlasov Control

2.1 Setup

Given a time horizon $T \leq \infty$, we define the MFCs, or MKV control problem with the tuple $(\mathcal{S},\ \mathcal{P}(\mathcal{S}),\ A,\ r,\ \gamma)$. Here $\mathcal{S}$ is the state space and A the action space, $\mathcal{P}(\mathcal{S})$ is the space of all probability measures on $\mathcal{S}$ and and $\mathcal{P}(A)$ the space of all probability measures on A. For example, if $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$, then $\mathcal{P}(\mathcal{S}) = \left\{(p_i)_{i=1}^{|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|} : \sum_{i=1}^{|\mathcal{S}|} p_i = 1, p_i \geq 0\right\}$. Note that $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(A)$ are always infinite dimensional unless $\mathcal{S}$ and A are finite.

In MFCs, a central controller dictates all agents. At each time t , t being a positive integer no more than T, the state of each agent is $s_t \in \mathcal{S}$. The central controller observes the probability distribution $\mu_t \in \mathcal{P}(\mathcal{S})$ of state s_t, (a.k.a. the population state distribution). Each agent takes an action $a_t \in A$ according to some policy π_t assigned by the central controller. Then each agent will receive a reward $r(s_t, \mu_t, a_t)$ and her state will move to the next state s_{t+1} according to a probability transition function of mean field type $P(s_t, \mu_t, \cdot)$. P and r are possibly unknown.

The accumulated reward of the central controller at the time t, given a random variable ξ and a policy $\boldsymbol{\pi}$, is defined as

$$v_t^{\boldsymbol{\pi}}(\xi) \quad = \quad \mathbb{E}^{\boldsymbol{\pi}}\left[\sum_{i=t}^{T} \gamma^{i-t} r_i | s_t = \xi\right], \tag{2.1}$$

subject to

$$s_{i+1} \sim P(s_i, \mu_i, a_i), \quad r_i = r(s_i, \mu_i, a_i), \quad a_i \sim \pi_i, \quad t \leq i \leq T,$$

where the factor $\gamma \in (0,1]$, $\mathbb{E}^{\boldsymbol{\pi}}$ denotes the expectation under policy $\boldsymbol{\pi}$, P: $\mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A \to \mathcal{P}(\mathcal{S})$ is probability transition function, and $r : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A \to \mathbb{R}$ immediate deterministic reward function. When $T < \infty$, we fix $\gamma = 1$. When $T = \infty$, we take $0 < \gamma < 1$.

The admissible control policies are restricted to be of a Markovian feedback form. That is, at each time $1 \le t \le T$, $\pi_t = \pi_t(s_t, \mu_t)$. Moreover, the control is a relaxed type, that is, $\pi_t : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(A)$ maps current state and the current state distribution to a distribution over the action space. This differs from a strict control α_t which is defined from $\mathcal{S} \times \mathcal{P}(\mathcal{S}) \to A$. In fact, a strict control is viewed as $\pi_t = \delta_{\alpha_t}$, the point mass at some α_t : $\mathcal{S} \times \mathcal{P}(\mathcal{S}) \to A$. We denote by Π_t the admissible policy set starting from time t. $\mathcal{A} = \{\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to A\}$ and $(\mathcal{P}(A))^{\mathcal{S}} = \{h : \mathcal{S} \to A\}$.

The goal of the central controller is to maximize over all policies the reward function

$$v_t(\xi) \;=\; \sup_{\boldsymbol{\pi} \in \Pi_t} v_t^{\boldsymbol{\pi}}(\xi), \quad 1 \le t \le T, \tag{2.2}$$

and to search for an optimal policy (if it exists) when P or r are unknown.

2.2 Preliminaries

Throughout the paper, we shall assume for the well-definedness of the problem

Outstanding Assumption (A).

$$\sup_{\boldsymbol{\pi} \in \Pi_1} \mathbb{E}^{\boldsymbol{\pi}} \big[\sum_{i=1}^{T} \gamma^{i-1} |r(s_i, \mu_i, a_i)| \big] \;<\; \infty.$$

Remark 2.1 *The following problem setting will ensure the Outstanding Assumption* (**A**). *Assume $\mathcal{S}$ and A are endowed with appropriate respective metrics $d_{\mathcal{S}}$ and d_A. Let $\mathcal{P}(\mathcal{S})$ be endowed with a Wasserstein distance of order 2 such that for any given two probability measures ν_1, $\nu_2 \in \mathcal{P}(\mathcal{S})$*

$$\mathcal{W}_2(\nu_1, \nu_2) = \left\{ \big(\inf \int_{\mathcal{S} \times \mathcal{S}} d_{\mathcal{S}}^2(x,y) \nu_{12}(dx, dy) \big)^{\frac{1}{2}} : \nu_{12} \in \mathcal{P}(\mathcal{S} \times \mathcal{S}) \text{ with marginals } \nu_1, \nu_2 \right\}.$$

If both $\mathcal{S}$ and A are finite, then we assume that $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(A)$ are compact and embedded in some Euclidean spaces. Moreover, assume

(A1) *For fixed arbitrary $(s^o, \delta_{s^o}, a^o) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$, δ_{s^o} being the dirac measure at point $s^o \in \mathcal{S}$, there exists some positive constant C such that for every $(s, \mu, a) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$*

$$\left| \int_{a \in A} d_A(a, a_0) (\pi(s, \mu, da) - \pi(s^o, \delta_{s^o}, da)) \right| \le C(1 + d_{\mathcal{S}}(s, s^o) + \mathcal{W}_2(\mu, \delta_{s^o})).$$

(A2) *For fixed arbitrary $(s^o, \delta_{s^o}, a^o) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$, there exists some positive constant C such that for every $(s, \mu, a) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$*

$$\begin{aligned} & \left| \int_{s' \in \mathcal{S}} d_{\mathcal{S}}^2(s', s^o) (P(s, \mu, a, ds') - P(s^o, \delta_{s^o}, a^o, ds')) \right| \\ \le \;& C \left(1 + d_{\mathcal{S}}^2(s, s^o) + d_A^2(a, a^o) + \mathcal{W}_2^2(\mu, \delta_{s^o}) \right). \end{aligned}$$

(A3) *For fixed arbitrary $(s^o, \delta_{s^o}, a^o) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$, there exists some positive constant C such that for every $(s, \mu, a) \in \mathcal{S} \times \mathcal{P}(\mathcal{S}) \times A$*

$$|r(s, \mu, a) - r(s^o, \delta_{s^o}, a^o)| \leq C\big(1 + d_{\mathcal{S}}^2(s, s^o) + d_A^2(a, a^o) + \mathcal{W}_2^2(\mu, \delta_{s^o})\big).$$

Note that classical MDP problems with finite state and action trivially satisfy Assumptions (A1)-(A2)-(A3). Note also when $\mathcal{S}$ or A is continuous space, one can use the norm instead of the metric in Assumption (A1)-(A2)-(A3). One can check that under Assumption (A1), for every policy $\boldsymbol{\pi} \in \Pi$, there exists some constant C with a bit use of notation

$$\mathbb{E}^{\boldsymbol{\pi}}[d_{\mathcal{S}}^2(s_{t+1}, s^o)] \leq C(1 + \mathbb{E}^{\boldsymbol{\pi}}[d_{\mathcal{S}}^2(s_t, s^o)]),$$

under Assumption (A2),

$$\mathbb{E}^{\boldsymbol{\pi}}[|r(s_t, \mu_t, a_t) - r(s^o, \delta_{s^o}, a^o)|] \leq C(1 + \mathbb{E}^{\boldsymbol{\pi}}[d_{\mathcal{S}}^2(s_t, s^o)]).$$

Then the Outstanding Assumption holds with suitable choice of γ with respect to the constant C.

Next, one can establish the following lemma for relaxed controls. This lemma shows that the value function $v_t^{\boldsymbol{\pi}}(\xi)$ can be rewritten in terms of state distribution flow $\boldsymbol{\mu} = \{\mu_t\}_t$ and depends on initial random variable ξ only through its probability distribution μ.

Lemma 2.1 *Given any $\boldsymbol{\pi} := \{\pi_i\}_{i=t}^T \in \Pi_t$, $v_t^{\boldsymbol{\pi}}(\xi)$ can be written in terms of μ_t. That is,*

$$v_t^{\boldsymbol{\pi}}(\mu) \;=\; \sum_{i=t}^{T} \gamma^{i-t} \hat{r}(\mu_i, \pi_i(\mu_i)), \tag{2.3}$$

where $\hat{r} : \mathcal{P}(\mathcal{S}) \times (\mathcal{P}(A))^{\mathcal{S}} \to \mathbb{R}$ is the integrated (averaged) reward function defined by

$$\hat{r}(\mu, h) \;:=\; \int_{s \in \mathcal{S}} \mu(ds) \int_{a \in A} h(s, da) r(s, \mu, a). \tag{2.4}$$

In particular, when $h = \delta_\alpha$ for some $\alpha : \mathcal{S} \to A$ (i.e., π_t is a strict control),

$$\hat{r}(\mu, h) \;=\; \int_{s \in \mathcal{S}} \mu(ds) r(s, \mu, \alpha(s, \mu)).$$

Proof. Lemma 2.1 is an immediate result of Fubini's theorem under Assumption (**A**).

$$\begin{aligned}
v_t^{\pi}(\xi) &= \mathbb{E}^{\pi}\Big[\sum_{i=t}^{T} \gamma^{i-t} r(s_i, \mu_i, a_i)\Big] \\
&= \mathbb{E}\Big[\sum_{i=t}^{T} \gamma^{i-t} \int_{a \in A} \pi_i(s_i, \mu_i, da) r(s_i, \mu_i, a)\Big] \\
&= \sum_{i=t}^{T} \gamma^{i-t} \mathbb{E}\Big[\int_{a \in A} \pi_i(s_i, \mu_i, da) r(s_i, \mu_i, a)\Big] \\
&= \sum_{i=t}^{T} \gamma^{i-t} \hat{r}(\mu_i, \pi_i(\mu_i)),
\end{aligned}$$

with $\hat{r}$ defined in (2.4).

We also have the flow property of μ_t. Note that the distribution flow $\{\mu_i\}_{i\geq t}$ depends on policy $\boldsymbol{\pi}$ and initial distribution μ starting from time t. In the rest of the paper, if needed, we shall stress this dependence by writing $\{\mu_i^{t,\mu,\boldsymbol{\pi}}\}_{i\geq t}$.

Lemma 2.2 *(Flow property of μ_t) Given any $h : \mathcal{S} \to \mathcal{P}(A)$, then the evolution of the state distribution is given by*

$$\mu_{t+1} = \int_{s\in\mathcal{S}} \mu(ds) \int_{a\in A} \pi_t(s_t, \mu_t, da) P(s_t, \mu_t, a, ds') := \Phi(\mu_t, \pi_t(\mu_t)), \tag{2.5}$$

Here

$$\Phi(\mu, h)(ds') := \int_{s\in\mathcal{S}} \mu(ds) \int_{a\in A} h(s, da) P(s, \mu, a, ds'). \tag{2.6}$$

In particular, when $h = \delta_\alpha$ for some $\alpha : \mathcal{S} \to A$ (i.e., π_t is a strict control),

$$\Phi(\mu, h)(ds') := \int_{s\in\mathcal{S}} \mu(ds) P(s, \mu, \alpha(s, \mu), ds').$$

Proof. Fix $\boldsymbol{\pi} \in \Pi_1$. For any bounded measurable function φ on $\mathcal{S}$, by the law of iterated conditional expectation:

$$\begin{aligned}
\mathbb{E}^{\boldsymbol{\pi}}[\varphi(s_{t+1})] &= \mathbb{E}^{\boldsymbol{\pi}}\Big[\mathbb{E}^{\boldsymbol{\pi}}\big[\varphi(s_{t+1})|s_1 \ldots, s_t\big]\Big] \\
&= \mathbb{E}^{\boldsymbol{\pi}}\Big[\int_{s'\in\mathcal{S}} \varphi(s') P(s_t, \mu_t, a_t, ds')\Big] \\
&= \int_{s'\in\mathcal{S}} \varphi(s') \mathbb{E}^{\boldsymbol{\pi}}\Big[P(s_t, \mu_t, a_t, ds')\Big] \\
&= \int_{s'\in\mathcal{S}} \varphi(s') \int_{s\in\mathcal{S}} \mu_t(ds) \int_{a\in A} \pi_t(s, \mu_t, da) P(s, \mu_t, a, ds').
\end{aligned}$$

□

Notice that the above argument holds for both a finite horizon and an infinite horizon case.

The above lemma on the flow property of μ is critical for establishing the DPP for learning MFCs. It suggests that MFCs may be viewed as an MDP problem with the state variable s_t replaced by the probability distribution flow μ_t.

3 Bellman equation of IQ function for learning MFCs

In this section, we will address the time inconsistency issue of Q-learning for MFC problem.

The first step is to define appropriate Q function for learning MFCs.

3.1 Learning MFC and time inconsistency: an example

Recall the classical RL for MDP problem: at each time $t = 1, \cdots, T$, the agent at state $s_t \in \mathcal{S}$ chooses her action $a_t \in A$ according to some policy $\pi : \mathcal{S} \to \mathcal{P}(A)$, she will then receive a reward $r(s_t, a_t)$ and her state moves to s_{t+1} according to $P(s_t, a_t, \cdot)$, where r and

P are possibly unknown. The Q function is used to derive the total reward for the agent given her current state s_t and action a_t:

$$Q_t(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s' \sim P(s_t, a_t, \cdot)} v_{t+1}(s_{t+1}),$$

where r is the reward function, v_t is the optimal value function of the problem at time t, and $P(s_t, a_t, \cdot)$ is the Markov transition probability.

However, such Q function update with the state variable s and the action variable a will not be time consistent for learning MFCs, as indicated below.

Example 3.1 *Take a two-state dynamic system with two choices of controls. The state space $\mathcal{S} = \{0, 1\}$, the action space $A = \{N, M\}$. Here the action N means to stay and M means to move. The dynamic $\{s_t\}_{t \geq 1}$ goes as follows: if the agent at time t is in state i (i.e., $s_t = i$) and if she takes the action $a_t = N$ according to the central controller's demand, then $s_{t+1} = i$; if she moves according to the central controller's demand, then $s_{t+1} = 1 - i$. ($i = 0, 1$). After each action she will receive a reward $-\mathcal{W}_2(\mu_t, B)$. Here μ_t denotes the probability distribution of the state at time t, and B is a given Binomial distribution with parameter p ($0 \leq p \leq 1$). Fix any arbitrary initial state distribution $\mu_1 = p_1 1_{\{s_1=1\}} + (1 - p_1) 1_{\{s_1=0\}}$.*

Assume now the central controller takes the standard form of Q function with the state variables s instead of the state distribution μ, then at each iteration t, the standard Q-learning update leads to

$$\begin{aligned} Q_{t+1}(s^1, s^2, a^1, a^2) \quad &= (1 - l) Q_t(s_1, s_2, a^1, a^2) \\ &\quad + l * (r_t + \gamma * \max_{(a^{1\prime}, a^{2\prime})} (Q_t(s^1, s^2, a^{1\prime}, a^{2\prime})). \end{aligned} \tag{3.7}$$

Here l is the learning rate, γ is the discount factor to balance the immediate and the future rewards, and r_t is the observed reward sampled from taking action (a_1, a_2). Here $(s^1, s^2) = (0, 1)$ is fixed from a central controller's perspective as she needs to coordinate players in both states.

Note that μ_t does not appear in this standard Q function.

Following this Q-update, the experiment result on the convergence of Q function with different initial actions is reported below, with $T = 100$, $\mu_1 = (0.3, 0.7)$, $p = 0.6$, $l = 0.4$ and $\gamma = 0.5$.

Table 1: Convergence of Q function with different initial actions.

Initial actions	$Q_T(s^1, s^2, N, N)$	$Q_T(s^1, s^2, N, M)$	$Q_T(s^1, s^2, M, N)$	$Q_T(s^1, s^2, M, M)$
$(a_1^1, a_1^2) = (N, N)$	-0.8	-0.83715584	-0.83715584	-0.83473101
$(a_1^1, a_1^2) = (N, M)$	-0.8	-0.82667008	-0.83715584	-0.83473101
$(a_1^1, a_1^2) = (M, N)$	-0.91068477	-0.8000001	-0.91569156	-0.87969395
$(a_1^1, a_1^2) = (M, M)$	-0.80000011	-0.89460224	-0.89460224	-0.87546019

Here $a_t^j \in \{N, M\}$ is the action from all players in state j at time $t \geq 1$.

Note the time inconsistency here: when the learner takes different initial actions, the Q table will converge to different values as $\{Q_T(s^1, s^2, i, j)\}_{i,j}$ shows. The culprit: with this Q function, the state space and the action space are not sufficiently rich to ensure the DPP or the Bellman optimality for (3.7).

3.2 IQ function for learning MFCs

Example 3.1 indicates the wrong form of the Q function for leaning MFCs. The question is, what is wrong with such Q function?

First, MFC or the MKV control problem is well recognized as a central controller's control problem: instead of maximizing reward for each individual agent, the objective in MFC is to maximize the collective reward from the perspective of the central controller. Now the central controller's value function should be dependent on μ the probability distribution of the state. Therefore, the Q function for MFC should be dependent on μ instead of s.

Secondly, Lemma 2.2 suggests that once a policy $\boldsymbol{\pi} \in \Pi_1$ is given, the dynamics of the state distribution is determined by $\mu_{t+1} = \Phi(\mu_t, \pi_t(\mu_t))$, which is a *deterministic process* through π_t in $\mathcal{P}(\mathcal{S})$. Therefore, the second argument of the Q function should be an element in $(\mathcal{P}(A))^{\mathcal{S}}$, rather than a single action in A or a probability distribution on A.

Hence, the proper definition for learning MFCs should take the following form.

Definition 3.1 (IQ for learning MFCs)

$$Q_t(\mu, h) \quad = \quad \sup_{\boldsymbol{\pi} \in \Pi_{t+1}} Q_t^{\boldsymbol{\pi}}(\mu, h), \tag{3.8}$$

where

$$Q_t^{\boldsymbol{\pi}}(\mu, h) \quad = \quad \mathbb{E}^{\boldsymbol{\pi}}\big[\sum_{i=t}^{T} \gamma^{i-t} r(s_i, \mu_i, a_i) | s_t \sim \mu, a_t \sim h(s_t)\big].$$

3.3 DPP: Bellman Equation for IQ function

Now we establish the DPP for this IQ function, in the form of the following Bellman equation.

Theorem 3.1 *(Bellman equation for IQ)*

(1) *Infinite horizon: for any $\mu \in \mathcal{P}(\mathcal{S})$ and $h \in (\mathcal{P}(A))^{\mathcal{S}}$, we have*

$$Q(\mu, h) \quad = \quad \hat{r}(\mu, h) + \gamma \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\Phi(\mu, h), h'). \tag{3.9}$$

(2) *Finite horizon: for any $(\mu, h) \in \mathcal{P}(\mathcal{S}) \times (\mathcal{P}(A))^{\mathcal{S}}$, we have*

$$Q_t(\mu, h) \quad = \quad \hat{r}(\mu, h) + \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} Q_{t+1}(\Phi(\mu, h), h'),$$

for $1 \leq t < T$, where $Q_T(\mu, h) = \hat{r}(\mu, h)$.

We prove for the infinite horizon case. The proof can be easily adapted to the finite time horizon case.

Two technical lemmas proceed the proof of Theorem 3.1. The first one shows that value function v and the IQ function are independent of time. The second one establishes the relation between the value function v and the IQ function, as in the classical MDP problem between the value function and the Q function.

Lemma 3.3 *Q_t in (3.8) and v_t in (2.2) are time-independent, i.e.,*

$$v_t(\mu) \;=\; \sup_{\boldsymbol{\pi}\in\Pi_1} v_1^{\boldsymbol{\pi}}(\mu), \quad Q_t(\mu,h) \;=\; \sup_{\boldsymbol{\pi}\in\Pi_2} Q_1^{\boldsymbol{\pi}}(\mu,h).$$

Here Π_t is the admissible policy set starting from time t.

Proof Let us introduce the pair of state distribution and policy $\{(\bar\mu_i,\bar\pi_i)\}_{i=1}^\infty$ by the shift

$$\bar\mu_i \;=\; \mu_{i+t-1}, \quad \bar\pi_i \;=\; \pi_{i+t-1}, \;\; i\in\mathbb{N}, \tag{3.10}$$

Given $\boldsymbol{\pi}=\{\pi_{i+t-1}\}_{i=1}^\infty$, and $\boldsymbol{\mu}=\{\mu_{i+t-1}\}_{i=1}^\infty$ starting from μ, from (2.6), we have by the construction of $\bar{\boldsymbol{\mu}}=\{\bar\mu_i\}_{i=1}^\infty$ and $\bar{\boldsymbol{\pi}}=\{\bar\mu_i\}_{i=1}^\infty$

$$\bar\mu_{i+1}^{1,\mu,\bar{\boldsymbol{\pi}}} = \Phi(\bar\mu_i^{1,\mu,\bar{\boldsymbol{\pi}}},\bar\pi_i(\bar\mu_i^{1,\mu,\bar{\boldsymbol{\pi}}})), \;\; i\in\mathbb{N}.$$

By (2.3),

$$v_t^{\boldsymbol{\pi}}(\mu) \;=\; \sum_{i=1}^\infty \gamma^{i-1}\hat r(\mu_{i+t-1},\pi_{i+t-1}(\mu_{i+t-1})) \;=\; \sum_{i=1}^\infty \gamma^{i-1}\hat r(\bar\mu_i,\bar\pi_i(\bar\mu_i)) = v_1^{\bar{\boldsymbol{\pi}}}(\mu).$$

Then, for any $\mu\in\mathcal{P}(\mathcal{S})$

$$v_t(\mu) \;=\; \sup_{\boldsymbol{\pi}\in\Pi_t} v_t^{\boldsymbol{\pi}}(\mu) \;=\; \sup_{\bar{\boldsymbol{\pi}}\in\Pi_1} \bar v_1^{\bar{\boldsymbol{\pi}}}(\mu),$$

Repeat the same argument for Q as for v

$$Q_t(\mu,h) \;=\; \sup_{\boldsymbol{\pi}\in\Pi_{t+1}} Q^{\boldsymbol{\pi}}(\mu,h) \;=\; \sup_{\boldsymbol{\pi}\in\Pi_2} Q_1^{\boldsymbol{\pi}}(\mu,h).$$

□

Therefore, in the remaining part of the paper, we shall omit t in $v_t(\mu)$ and $Q_t(\mu,h)$ in the infinite horizon case.

Lemma 3.4 *For any $\mu\in\mathcal{P}(\mathcal{S})$,*

$$v(\mu) \;=\; \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} Q(\mu,h). \tag{3.11}$$

Proof. Fix some arbitrary $\mu\in\mathcal{P}(\mathcal{S})$ and any given $\boldsymbol{\pi}=\{\pi_i\}_{i=t}^\infty\in\Pi_t$, we have

$$\begin{aligned} v_t^{\boldsymbol{\pi}}(\mu) &= \hat r(\mu,\pi_t(\mu)) + \sum_{i=t+1}^\infty \gamma^{i-t}\hat r(\mu_i,\pi_i(\mu_i)) \\ &= \hat r(\mu,\pi_t(\mu)) + \gamma v_{t+1}^{\boldsymbol{\pi}_{-t}}(\Phi(\mu,\pi_t(\mu))), \\ &= Q_t^{\boldsymbol{\pi}_{-t}}(\mu,\pi_t(\mu)), \end{aligned} \tag{3.12}$$

where we denote by $\boldsymbol{\pi}_{-t} := \{\pi_i\}_{i=t+1}^\infty$, and in the second equality, we used the Markov property of $\{\mu_i^{t,\mu,\boldsymbol{\pi}}\}_{i=t}^\infty$ from Lemma 2.2.

To prove (3.11). First, we show $v(\mu)\le\sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} Q(\mu,h)$. To see this, note

$$v_t^{\boldsymbol{\pi}}(\mu) \;=\; Q_t^{\boldsymbol{\pi}_{-t}}(\mu,\pi_t(\mu)) \;\le\; Q(\mu,\pi_t(\mu)) \;\le\; \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} Q(\mu,h), \tag{3.13}$$

where the first inequality is by definition of $Q(\mu, h)$, and $\pi_t(\mu) \in (\mathcal{P}(A))^{\mathcal{S}}$ for each $\mu \in \mathcal{P}(\mathcal{S})$. Taking supremum over all policies $\boldsymbol{\pi} \in \Pi_t$ in (3.13) shows that

$$v(\mu) \quad \leq \quad \sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\mu, h). \tag{3.14}$$

To see $v(\mu) \geq \sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\mu, h)$, fix any arbitrary $\mu \in \mathcal{P}(\mathcal{S})$, $\pi_t \in \mathcal{A}$, for any $\epsilon > 0$, there exists $\boldsymbol{\pi}^{\epsilon} = \{\pi_l^{\epsilon}\}_{l=t+1}^{\infty}$ such that

$$Q_t^{\boldsymbol{\pi}^{\epsilon}}(\mu, \pi_t(\mu)) \geq Q(\mu, \pi_t(\mu)) - \epsilon. \tag{3.15}$$

Now define $\tilde{\boldsymbol{\pi}} = \{\tilde{\pi}_l\}_{l=t}^{\infty} \in \Pi_t$ by

$$\tilde{\pi}_i \quad = \quad \pi_t 1_{\{i=t\}} + \pi_i^{\epsilon} 1_{\{i=t+1,\ldots\}},$$

then from (3.12) and (3.15)

$$v(\mu) \quad \geq \quad v_t^{\tilde{\boldsymbol{\pi}}}(\mu) \; = \; Q_t^{\boldsymbol{\pi}^{\epsilon}}(\mu, \pi_t(\mu)) \geq Q(\mu, \pi_t(\mu)) - \epsilon, \tag{3.16}$$

Taking supremum over all $\pi_t \in \mathcal{A}$ in (3.16), we obtain

$$v(\mu) \quad \geq \quad \sup_{\pi \in \mathcal{A}} Q(\mu, \pi(\mu)) - \epsilon \; = \; \sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\mu, h) - \epsilon.$$

Since the above inequality holds for any $\epsilon > 0$,

$$v(\mu) \quad \geq \quad \sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\mu, h). \tag{3.17}$$

(3.11) follows from (3.14) and (3.17). □

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1.

Note that $Q(\mu, h) = \sup_{\pi \in \Pi_{t+1}} Q_t^{\pi}(\mu, h)$ for any $t \in \mathbb{N}$. Without loss of generality take $t = 1$, then

$$\begin{aligned} Q(\mu, h) \quad &= \quad \sup_{\boldsymbol{\pi} \in \Pi_2} \mathbb{E}^{\boldsymbol{\pi}} \big[\sum_{i=1}^{\infty} \gamma^{i-1} r(s_i, \mu_i, a_i) | s_1 \sim \mu, a_1 \sim h(s_1) \big] \\ &= \quad \sup_{\boldsymbol{\pi} \in \Pi_2} \big[\hat{r}(\mu, h) + \gamma v_2^{\boldsymbol{\pi}}(\Phi(\mu, h)) \big] \\ &= \quad \hat{r}(\mu, h) + \gamma v(\Phi(\mu, h)) \\ &= \quad \hat{r}(\mu, h) + \gamma \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} Q(\Phi(\mu, h), h'), \end{aligned}$$

where the second equality is from the Markov property of μ_t, $t \in \mathbb{N}$, the third equality is by the definition of the value function, and the last inequality is from Lemma 3.4. □

So far, we have established the necessary condition for the Bellman optimality. That is, the IQ function satisfies the Bellman equation and is time consistent. We can further establish that this Bellman equation is sufficient, as the following verification theorem.

Proposition 3.1 (Verification theorem) *If $\tilde{Q} : \mathcal{P}(\mathcal{S}) \times (\mathcal{P}(A))^{\mathcal{S}} \to \mathbb{R}$ satisfies Bellman relation* (3.9) *with $\lim_{t \to \infty} \gamma^t \tilde{Q}(\mu, h) = 0$ for any $(\mu, h) \in \mathcal{P}(\mathcal{S}) \times (\mathcal{P}(A))^{\mathcal{S}}$. Suppose that for every $\mu \in \mathcal{P}(\mathcal{S})$, one can also find a stationary policy $\pi^*(\mu) \in (\mathcal{P}(A))^{\mathcal{S}}$ that achieves $\sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\mu, h)$, then π^* is an optimal policy of problem* (2.2).

Proof. On one hand, given any $\mu \in \mathcal{P}(\mathcal{S})$, for any given policy $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^{\infty} \in \Pi_1$, the evolution of $\{\mu_i^{1,\mu,\boldsymbol{\pi}}\}_{i=1}^{\infty}$ is given by(2.5). From (3.9)

$$\tilde{Q}(\mu_i, \pi_i(\mu_i)) \quad \geq \quad \hat{r}(\mu_i, \pi_i(\mu_i)) + \gamma \tilde{Q}(\mu_{i+1}, \pi_{i+1}(\mu_{i+1})), \quad i \in \mathbb{N},$$

multiplying by γ^{i-1} and rearranging yield

$$\gamma^{i-1}\tilde{Q}(\mu_i, \pi_i(\mu_i)) - \gamma^i \tilde{Q}(\mu_{i+1}, \pi_i(\mu_i)) \quad \geq \quad \hat{r}(\mu_i, \pi_i(\mu_i)), \quad i \in \mathbb{N},$$

Taking summation over $1 \leq i \leq t-1$, we obtain

$$\tilde{Q}(\mu, \pi_1(\mu)) - \gamma^{t-1}\tilde{Q}(\mu_t, \pi_l(\mu_t)) \quad \geq \quad \sum_{i=1}^{t-1} \gamma^{i-1} \hat{r}(\mu_i, \pi_i(\mu_i)), \quad 1 \leq i \leq t-1.$$

Given $\lim_{t\to\infty} \gamma^t \tilde{Q}(\mu, h) = 0$, by taking the limit $t \to \infty$,

$$\tilde{Q}(\mu, \pi_1(\mu)) \quad \geq \quad \sum_{i=1}^{\infty} \gamma^{i-1} \hat{r}(\mu_i, \pi_i(\mu_i)) = v_1^{\boldsymbol{\pi}}(\mu).$$

Therefore,

$$\sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\mu, h) \geq v(\mu).$$

On the other hand, since $\pi^*(\mu) \in \arg\max \tilde{Q}(\mu, h)$ for every $\mu \in \mathcal{P}(\mathcal{S})$, then for every $i \geq 1$

$$\tilde{Q}(\mu_i^{1,\mu,\pi^*}, \pi^*(\mu_i^{1,\mu,\pi^*})) \quad = \quad \hat{r}(\mu_i^{1,\mu,\pi^*}, \pi^*(\mu_i^{1,\mu,\pi^*})) + \gamma \tilde{Q}(\mu_{i+1}^{1,\mu,\pi^*}, \pi^*(\mu_{i+1}^{1,\mu,\pi^*})).$$

Repeat the same argument for π^* as for π

$$\sup_{h \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\mu, h) = \tilde{Q}(\mu, \pi^*(\mu)) = v_1^{\pi^*}(\mu),$$

which shows that π^* is an optimal policy. □

3.4 IQ function vs classical Q function.

As discussed earlier, the appropriate IQ function for learning MFCs is to "lift" the classical Q function for learning MDPs by replacing the state space $\mathcal{S}$ and action space A with the state space $\mathcal{P}(\mathcal{S})$ and action space $(\mathcal{P}(A))^{\mathcal{S}}$ respectively. There is a more precise and analytical connection between their respective Bellman equations, and hence the term of *IQ* function.

To see this, recall the classical Q-learning for an infinite time horizon MDP problem, when there is no state distribution in the probability transition function P or in the reward function r. For simplicity, we assume $\mathcal{S}$ and A are finite space so that r is bounded. Then Q in (3.9) is the integral of $\bar{Q}$ of the following form,

$$Q(\mu, h) \quad = \quad \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} \bar{Q}(s, a) h(s, a) \tag{3.18}$$

where $\bar{Q}(s, a)$ satisfies the Bellman equation for standard MDP:

$$\bar{Q}(s, a) \quad = \quad r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') \max_{a' \in A} \bar{Q}(s', a'). \tag{3.19}$$

To derive this connection more precisely, define

$$\tilde{Q}(\mu, h) = \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} \bar{Q}(s, a) h(s, a).$$

Note that $\tilde{Q}$ is linear in μ and h. From (3.19), we have

$$\tilde{Q}(\mu, h) \;=\; \hat{r}(\mu, h) + \gamma \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \max_{a' \in A} \bar{Q}(s', a'),$$

then we can see that

$$\sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \max_{a' \in A} \bar{Q}(s', a') \;=\; \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\Phi(\mu, h), h'). \quad (3.20)$$

In fact, on one hand, for any $h' \in (\mathcal{P}(A))^{\mathcal{S}}$,

$$\begin{aligned}
&\sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \max_{a' \in A} \bar{Q}(s', a') \\
&= \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \sum_{\tilde{a} \in A} h'(s', \tilde{a}) \max_{a' \in A} \bar{Q}(s', a') \\
&\geq \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \sum_{\tilde{a} \in A} h'(s', \tilde{a}) \bar{Q}(s', \tilde{a}) \\
&= \sum_{s' \in \mathcal{S}} \Phi(\mu, h)(s') \sum_{\tilde{a} \in A} h'(s', \tilde{a}) \bar{Q}(s', \tilde{a}) \\
&= \tilde{Q}(\Phi(\mu, h), h'),
\end{aligned}$$

where the first equality is from $\sum_{\tilde{a} \in A} h(s', \tilde{a}) = 1$, the second equality is from (2.6), and the last equality is from the definition of $\tilde{Q}$.

On the other hand, if we take

$$h'_*(s') \;=\; \begin{cases} 1, & a_*(s') \in \arg\max_{a' \in A} \bar{Q}(s', a') \\ 0, & \text{otherwise,} \end{cases}$$

then

$$\begin{aligned}
&\sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\Phi(\mu, h), h') \\
&\geq \tilde{Q}(\Phi(\mu, h), h'_*) \\
&= \sum_{s' \in \mathcal{S}} \Phi(\mu, h)(s') \sum_{a' \in A} \bar{Q}(s', a') h'_*(s', a') \\
&= \sum_{s' \in \mathcal{S}} \Phi(\mu, h)(s') \max_{a' \in A} \bar{Q}(s', a') \\
&= \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in A} h(s, a) \sum_{s' \in \mathcal{S}} P(s, a, s') \max_{a' \in A} \bar{Q}(s', a').
\end{aligned}$$

Therefore,

$$\tilde{Q}(\mu, h) \;=\; \hat{r}(\mu, h) + \gamma \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} \tilde{Q}(\Phi(\mu, h), h').$$

Since both $\tilde{Q}$ and Q satisfy Bellman equations (3.9), we have $Q = \tilde{Q}$ from the uniqueness of the fixed point of a contraction mapping $(Fw)(\mu, h) = \hat{r}(\mu, h) + \gamma \sup_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} w(\Phi(\mu, h), h')$.

3.5 DPP for the value function

Q-learning update is essential for model-free RL of control problems, while model- based learning algorithms such as the value iteration rely on the DPP for the value function. We see in fact, the Bellman equation for v is a simple corollary of Lemma 3.4 and Theorem 3.1.

Corollary 3.1 (DPP for value function) *The value function v satisfies the Bellman equation*

$$v(\mu) \quad = \quad \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \big[\hat{r}(\mu,h) + \gamma v(\Phi(\mu,h))\big], \tag{3.21}$$

for any $\mu \in \mathcal{P}(\mathcal{S})$.

Now, with this DPP for the value function, we can design the value iteration. Given $\pi \in \mathcal{A}$, define the operator $T_\pi : \mathcal{B}(\mathcal{P}(\mathcal{S})) \to \mathcal{B}(\mathcal{P}(\mathcal{S}))$ such that

$$(T_\pi w)(\mu) \quad := \quad \hat{r}(\mu,\pi(\mu)) + \gamma w(\Phi(\mu,\pi(\mu))), \tag{3.22}$$

and another operator $T : \mathcal{B}(\mathcal{P}(\mathcal{S})) \to \mathcal{B}(\mathcal{P}(\mathcal{S}))$ such that

$$(Tw)(\mu) \quad := \quad \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \big[\hat{r}(\mu,h) + \gamma w(\Phi(\mu,h))\big], \tag{3.23}$$

where $\mathcal{B}(\mathcal{P}(\mathcal{S}))$ is the set of all measurable functions on $\mathcal{P}(\mathcal{S})$, $\hat{r}(\mu,h)$ and $\Phi(\mu,h)$ are given in (2.4) and (2.6).

Proposition 3.2 *Assume without loss of generality $v_1 = 0$, then under Assumption (**A**), we have for all $\mu \in \mathcal{P}(\mathcal{S})$,*

$$v(\mu) \quad = \quad \lim_{n\to\infty} (T^n v_1)(\mu).$$

where T^n is n times composition of T.

Proof of Proposition 3.2 relies on the following Lemma.

Lemma 3.5 *Assume without loss of generality $v_1 = 0$, then for any $\mu \in \mathcal{P}(\mathcal{S})$ and $\boldsymbol{\pi} = \{\pi_t\}_{t=1}^n$ with $\pi_t \in \mathcal{A}$ for every $1 \le t \le n$,*

$$(T_{\pi_1}\cdots T_{\pi_n} v_1)(\mu) \quad = \quad \sum_{t=1}^{n} \gamma^t \hat{r}(\mu_t, \pi_t(\mu_t)), \tag{3.24}$$

$$(T^n v_1)(\mu) \quad = \quad \sup_{\{\pi_t\}_{t=1}^n} (T_{\pi_1}\ldots T_{\pi_n} v_1)(\mu), \tag{3.25}$$

where $T_{\pi_1}\cdots T_{\pi_n}$ is the composition of all T_{π_i}.

Proof. The proof of (3.24) and (3.25) are by the forward induction. Here we state the proof of (3.25). The result clearly holds for $n = 1$ as

$$\sup_{\pi_1\in\mathcal{A}} (T_{\pi_1} v_1)(\mu) = \sup_{\pi_1\in\mathcal{A}} \hat{r}(\mu,\pi_1(\mu)) = \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \hat{r}(\mu,h) = (Tv_1)(\mu).$$

Suppose that (3.25) holds for $n = k$. Then when $n = k + 1$

$$\begin{aligned}
(T^{k+1}v_1)(\mu) &= \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \left[\hat{r}(\mu,h) + \gamma(T^k v_1)(\Phi(\mu,h))\right] \\
&= \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \left[\hat{r}(\mu,h) + \gamma \sup_{\{\tilde{\pi}_t\}_{t=1}^{k}} (T_{\tilde{\pi}_1}\dots T_{\tilde{\pi}_k}v_1)(\Phi(\mu,h))\right] \\
&= \sup_{h\in(\mathcal{P}(A))^{\mathcal{S}}} \left[\hat{r}(\mu,h) + \gamma \sup_{\{\pi_t\}_{t=2}^{k+1}} (T_{\pi_2}\dots T_{\pi_{k+1}}v_0)(\Phi(\mu,h))\right] \\
&= \sup_{h,\{\pi_t\}_{t=2}^{k+1}} \left[\hat{r}(\mu,h) + \gamma(T_{\pi_2}\dots T_{\pi_{k+1}}v_1)(\Phi(\mu,h))\right] \\
&= \sup_{\{\pi_t\}_{t=1}^{k+1}} (T_{\pi_1}\dots T_{\pi_{k+1}}v_1)(\mu),
\end{aligned}$$

where the first equality is from the definition of T in (3.25); the second equality is by the assumption that (3.25) holds at time $n = k$. □

Proof of Proposition 3.2 We write $v^{\boldsymbol{\pi}}(\mu)$ in two parts

$$\begin{aligned}
v^{\boldsymbol{\pi}}(\mu) &= \sum_{t=1}^{n} \gamma^t \hat{r}(\mu_t, \pi_t(\mu_t)) + \sum_{t=n+1}^{\infty} \gamma^t \hat{r}(\mu_t, \pi_t(\mu_t)) \\
&= (T_{\pi_1}\dots T_{\pi_n}v_1)(\mu) + \sum_{t=n+1}^{\infty} \gamma^t \hat{r}(\mu_t, \pi_t(\mu_t)), \qquad (3.26)
\end{aligned}$$

where the second equality is by (3.24).

Assumption (**A**) implies $\lim_{n\to\infty} \sup_{\boldsymbol{\pi}} \sum_{t=n+1}^{\infty} \gamma^t |\hat{r}(\mu_t, \pi_t(\mu_t))| = 0$. Taking supremum over $\boldsymbol{\pi} \in \Pi_1$ in (3.26) together with (3.25) gives

$$v(\mu) \leq (T^n v_1)(\mu) + \gamma \sup_{\boldsymbol{\pi}} \sum_{t=n+1}^{\infty} \gamma^t |\hat{r}(\mu_t^{1,\mu,\boldsymbol{\pi}}, \pi_t(\mu_t^{1,\mu,\boldsymbol{\pi}}))|, \qquad (3.27)$$

$$v(\mu) \geq (T^n v_1)(\mu) - \gamma \sup_{\boldsymbol{\pi}} \sum_{t=n+1}^{\infty} \gamma^t |\hat{r}(\mu_t^{1,\mu,\boldsymbol{\pi}}, \pi_t(\mu_t^{1,\mu,\boldsymbol{\pi}}))|. \qquad (3.28)$$

Taking the limit as $n \to \infty$ together with (3.25) yields

$$v(\mu) = \lim_{n\to\infty} (T^n v_1)(\mu).$$

□

4 Example 3.1 revisited

We now revisit Example 3.1. We first show that the optimal control is indeed a relaxed type. We then illustrate through numerical experiment the time consistency with the IQ function.

Example 4.2 (Example 3.1 revisited.) *Take a two-state dynamic system with two choices of controls. The state space $\mathcal{S} = \{0, 1\}$, the action space $A = \{N, M\}$. Here the action N means stay and M means move. The dynamic $(s_t)_{t\geq 0}$ goes as follows: if the agent at time t is in state i (i.e., $s_t = i$) and if she takes the action $a_t = N$, then $s_{t+1} = i$; if she decides*

to move then $s_{t+1} = 1 - i$. ($i = 0, 1$). In each round t, after actions from each individual, the central controller will receive a reward $-\mathcal{W}_2(\mu_t, B)$. Here μ_t denotes the probability distribution of the state at time t, and B is a given Binomial distribution with parameter p ($0 \le p \le 1$). Fix any arbitrary initial state distribution $\mu_1 = p_1 1_{\{s_1=1\}} + (1-p_1)1_{\{s_1=0\}}$.

Note that $-\mathcal{W}_2(\mu_t, B) \le 0$, we have for each policy $\boldsymbol{\pi} = \{\pi_l\}_{l=1}^{\infty} \in \Pi_1$,

$$v^{\boldsymbol{\pi}}(\mu_0) \;=\; -\sum_{t=1}^{\infty} \gamma^t \mathcal{W}_2(\mu_t, B) \;\le\; -\mathcal{W}_2(\mu_1, B),$$

the equality holds, i.e., the optimal value function is attained, if and only if the dynamic of state distribution corresponding to optimal policy π^* is given by

$$\mu_t \;=\; B, \;\; t \ge 2, \quad \mu_1 \;=\; p_1 1_{\{s=1\}} + q_1 1_{\{s=2\}}.$$

From (2.5), we get

$$\Phi(\mu, h) \;=\; \Big(\mu(1)h(1,S) + \mu(2)h(2,M), 1 - \mu(1)h(1,S) - \mu(2)h(2,M)\Big), \quad (4.29)$$

hence

$$\begin{aligned} p \;=\; \mu_2(1) \;&=\; \mu_1(1)\pi^*(1,S) + \mu_1(2)\pi^*(2,M) \;=\; p_1\pi^*(1,S) + q_1\pi^*(2,M), \\ p \;=\; \mu_{t+1}(1) \;&=\; \mu_t(1)\pi^*(1,S) + \mu_t(2)\pi^*(2,M) \;=\; p\pi^*(S|1) + q\pi^*(2,M), \; t \ge 2, \end{aligned}$$

which gives a stationary optimal policy

$$\pi^*(1) \;=\; p1_{\{a=S\}} + q1_{\{a=M\}}, \quad \pi^*(2) \;=\; q1_{\{a=S\}} + p1_{\{a=M\}}. \quad (4.30)$$

Now, the Q-learning update at each iteration t using the IQ function is

$$Q_{t+1}(\mu, h) \;=\; Q_t(\mu, h) + l * \left(\hat{r}_t + \gamma * \left(\max_{h' \in (\mathcal{P}(A))^{\mathcal{S}}} Q_t(\mu', h') - Q_t(\mu, h)\right)\right), \quad (4.31)$$

Here l is the learning rate and γ is the discount factor. In the algorithm, we shall use element $(p, 1-p)$ in the Euclidean space $\mathbb{R}^2$ to denote the Binomial distribution with parameter p.

Next, we design a simple algorithm (Algorithm 1) to show the performance of the IQ update (4.31), with the following specifications.

(a) **Dimension reduction**: Since $\mu_t(1)+\mu_t(2) = 1$ ($t = 1, 2, \cdots, T$) , $\pi(1,S)+\pi(1,M) = 1$ and $\pi(2,S) + \pi(2,M) = 1$ for any distribution μ_t and policy π, we can reduce the dimension of the IQ function. If we define $Q(\mu^1, \alpha_S^1, \alpha_S^2)$, with μ^1 the population probability at state 1, α_S^1 the action probability to "stay" at state 1, and α_S^2 the action probability to "stay" at state 2, then $Q(\mu^1, \alpha_S^1, \alpha_S^2) = Q(\mu^1, \mu^2, \alpha_S^1, \alpha_M^1, \alpha_S^2, \alpha_M^2)$.

(b) **Distribution discretization**: Classic Q-learning algorithms are designed for discrete state and action spaces. To examine the time-consistency property of (3.9) we discretize the state and action distribution with finite precision and apply the classic Q-learning update to (4.32) with finite-dimensional inputs. For simplicity, we assume uniform discretization such that $\tilde{P}(A) := \{i/N_a : 0 \le i \le N_a\}$ and $\tilde{P}(S) := \{i/N_s : 0 \le i \le N_s\}$ for some constant integers $N_a > 0$ and $N_s > 0$. For more refined discretization beyond the uniform one, see ϵ-Net in [14]. Note here the focus is to test the time consistency instead of designing the most efficient algorithm.

(c) **Algorithmic design**: The algorithm is summarized in Algorithm 1. Note that (4.32) is the reduced form of the original update (4.31) with discretized distribution. In order to perform the for-loop (Step 3, 4, and 5) in Algorithm 1, we assume the accessibility to a population simulator $(\mu', \hat{r}) = \mathcal{G}(\mu, \pi)$.

(d) **Metric design**: Explicit calculations show that the stationary optimal policy is given by (4.30). Therefore, we design the following matrix to check the convergence of the Q table to the true value and the speed of the convergence.

$$E(t) = \frac{1}{N_s} \sum_{i=0}^{N_s} \left| Q_t \left(\frac{i}{N}, p, q \right) + \mathcal{W}_2 \left(\left(\frac{i}{N}, 1 - \frac{i}{N} \right), B \right) \right|.$$

Here for simplicity we take $N_s = N_a = N$.

(e) **Parameter set-up**: Parameters are set as follows: $T = 15$, $p = 0.6$, $lr = 0.4$, $\gamma = 0.5$, and $N_a = N_s = 10$. Each component in Q_0 is randomly initialized from a uniform distribution on $[0, 1]$. The experiments are repeated 20 times.

(f) **Performance analysis.** The experiments shows that matrix $E(t)$ converges in around 10 outer iterations (Figure 1). The standard deviation of 20 repeated experiments is very small. This is partially due to the deterministic property of the underlying system.

Recall $\tilde{P}(\mathcal{S}) = \{\frac{i}{N_s} : 0 \le i \le N_s\}$. Further denote the projection as

$$\text{Proj}(\Phi^1(\mu^1, \alpha_S^1, \alpha_S^2), \tilde{P}(\mathcal{S})) := \text{argmin}_{\tilde{\mu}^1 \in \tilde{P}(\mathcal{S})} |\Phi^1(\mu^1, \alpha_S^1, \alpha_S^2) - \tilde{\mu}^1|.$$

Then the algorithm is summarized as follows.

Algorithm 1 MFC Q-learning with distribution discretization

1: **Input**: N_a and N_s.
2: **for** $t = 1, \cdots, T$ **do**
3: **for** $\alpha_S^1 \in \{\frac{i}{N_a}, 0 \le i \le N_a\}$ **do**
4: **for** $\alpha_S^2 \in \{\frac{i}{N_a}, 0 \le i \le N_a\}$ **do**
5: **for** $\mu^1 \in \{\frac{i}{N_s}, 0 \le i \le N_s\}$ **do**
6: $\mu^{1\prime} = \text{Proj}(\Phi^1(\mu^1, \alpha_S^1, \alpha_S^2), \tilde{P}(\mathcal{S}))$
7:

$$\begin{aligned} Q_{t+1}(\mu^1, \alpha_S^1, \alpha_S^2) &= (1-l) Q_t(\mu^1, \alpha_S^1, \alpha_S^2) \\ &\quad + l * \left(\hat{r}_t + \gamma * \max_{(\alpha_S^{1\prime}, \alpha_S^{2\prime}) \in (\tilde{P}(A))^2} Q_t(\mu_1', \alpha_S^{1\prime}, \alpha_S^{2\prime}) \right), \end{aligned} \tag{4.32}$$

8: **end for**
9: **end for**
10: **end for**
11: **end for**

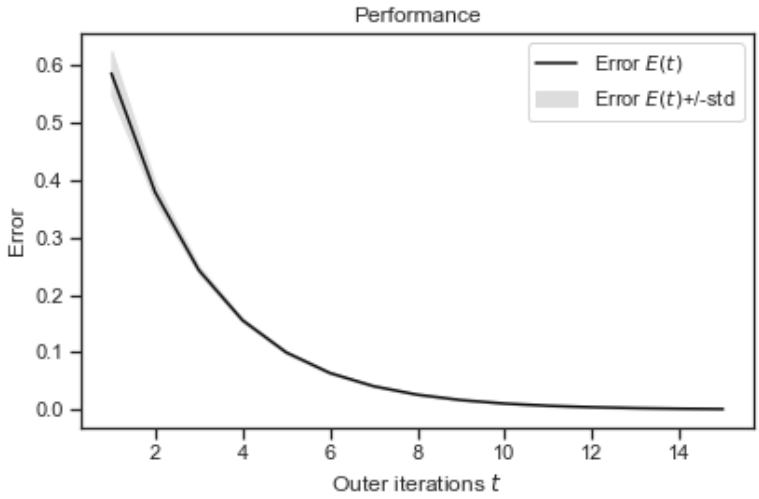
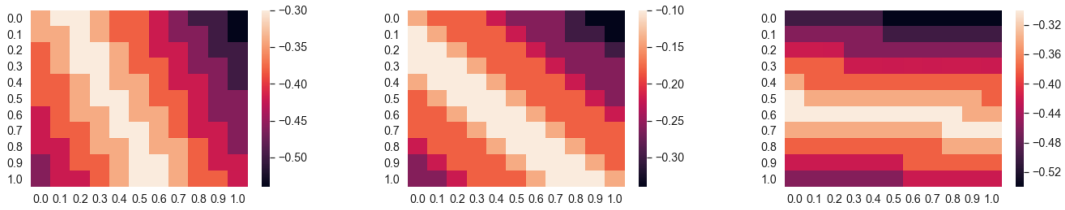


Figure 1: Numerical Performance on IQ Iterations.



(a) $Q_T(0.3,\cdot,\cdot)$ (b) $Q_T(0.5,\cdot,\cdot)$. (c) $Q_T(0.9,\cdot,\cdot)$

Figure 2: Snapshots of the IQ tables at final iteration T.

Remark 4.2 *In general, distribution discretization is sample-inefficient and suffers from the curse of dimensionality. For example, in Example* (3.1)*, there are two states and two actions, with* $N = N_s = N_a = 10$ *with precision* 0.1*. The Q function is a table of dimension* 1000*. This complexity grows exponentially with the number of states and actions. Moreover, although* $E(t)$ *converges relatively fast, there is unavoidable errors due to truncation, as seen in Figure 2. The optimal value* $Q(\frac{i}{N}, p, q)$ *can not be distinguished from its surrounding areas, where the areas with the lightest color all correspond to the largest value. This is because the accuracy is only up to* 0.1 *in each iteration. Therefore, it is desirable to develop sample-efficient and accurate Q-learning algorithms for learning MFC with the correct Bellman update* (3.9)*. This is our next research project.*

References

[1] René Aïd, Matteo Basei, and Huyên Pham. The coordination of centralised and distributed generation. *arXiv preprint arXiv:1705.01302*, 2017.

[2] Daniel Andersson and Boualem Djehiche. A maximum principle for SDEs of mean-field type. *Applied Mathematics & Optimization*, 63(3):341–356, 2011.

[3] Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[4] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

[5] Rainer Buckdahn, Boualem Djehiche, and Juan Li. A general stochastic maximum principle for SDEs of mean-field type. *Applied Mathematics & Optimization*, 64(2):197–216, 2011.

[6] René Carmona and François Delarue. Forward–backward stochastic differential equations and controlled McKean–Vlasov dynamics. *The Annals of Probability*, 43(5):2647–2700, 2015.

[7] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

[8] René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[9] René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[10] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *Aaai/iaai*, pages 761–768, 1998.

[11] Mao Fabrice Djete, Dylan Possamaï, and Xiaolu Tan. McKean-Vlasov optimal control: the dynamic programming principle. *arXiv preprint arXiv:1907.08860*, 2019.

[12] Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, 2000.

[13] Josselin Garnier, George Papanicolaou, and Tzu-Wei Yang. Large deviations for a mean field model of systemic risk. *SIAM Journal on Financial Mathematics*, 4(1):151–184, 2013.

[14] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *arXiv preprint arXiv:1901.09585*, 2019.

[15] Juho Hamari, Mimmi Sjöklint, and Antti Ukkonen. The sharing economy: Why people participate in collaborative consumption. *Journal of the association for information science and technology*, 67(9):2047–2059, 2016.

[16] Minyi Huang, Roland P Malhamé, Peter E Caines, et al. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

[17] Seong Hoon Jeong, Ah Reum Kang, and Huy Kang Kim. Analysis of game bot's behavioral characteristics in social interaction networks of MMORPG. In *ACM SIGCOMM Computer Communication Review*, volume 45, pages 99–100. ACM, 2015.

[18] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

[19] Daniel Lacker. Limit theory for controlled McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(3):1641–1672, 2017.

[20] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[21] Mathieu Laurière and Olivier Pironneau. Dynamic programming for mean-field type control. *Comptes Rendus Mathematique*, 352(9):707–713, 2014.

[22] Charles-Albert Lehalle and Charafeddine Mouzouni. A mean field game of portfolio trading and its consequences on perceived correlations. *arXiv preprint arXiv:1902.09606*, 2019.

[23] Shie Mannor and John N Tsitsiklis. Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.

[24] H McKean. Propagation of chaos for a class of non-linear parabolic equations. lecture series in differential equations 7. *Stochastic Differential Equations*, pages 41–57, 1969.

[25] Galo Nuño. Optimal social policies in mean field games. *Applied Mathematics & Optimization*, 76(1):29–57, 2017.

[26] Huyên Pham and Xiaoli Wei. Discrete time McKean–Vlasov control problem: a dynamic programming approach. *Applied Mathematics & Optimization*, 74(3):487–506, 2016.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Haoran Wang and Xun Yu Zhou. Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Available at SSRN 3382932*, 2019.

[29] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: A reinforcement learning approach to mean-field control via mean embedding. *Manuscript*, 2019.

[30] Christopher JCH Watkins. Learning from delayed rewards. *PhD thesis*, 1989.

[31] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[32] Michael Wunder, Michael L Littman, and Monica Babes. Classes of multiagent Q-learning dynamics with epsilon-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1167–1174. Citeseer, 2010.

[33] Xun Yu Zhou. On the existence of optimal relaxed controls of stochastic partial differential equations. *SIAM journal on control and optimization*, 30(2):247–261, 1992.