



**HAL**  
open science

# Mean-field Markov decision processes with common noise and open-loop controls

Médéric Motte, Huyên Pham

► **To cite this version:**

Médéric Motte, Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. 2021. hal-02413114v2

**HAL Id: hal-02413114**

**<https://hal.science/hal-02413114v2>**

Preprint submitted on 8 Sep 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Mean-field Markov decision processes with common noise and open-loop controls

Médéric MOTTE \*      Huyên PHAM †

This version: june 2021  
to appear in *Annals of Applied Probability*

## Abstract

We develop an exhaustive study of Markov decision process (MDP) under mean field interaction both on states and actions in the presence of common noise, and when optimization is performed over open-loop controls on infinite horizon. Such model, called CMKV-MDP for conditional McKean-Vlasov MDP, arises and is obtained here rigorously with a rate of convergence as the asymptotic problem of  $N$ -cooperative agents controlled by a social planner/influencer that observes the environment noises but not necessarily the individual states of the agents. We highlight the crucial role of relaxed controls and randomization hypothesis for this class of models with respect to classical MDP theory. We prove the correspondence between CMKV-MDP and a general lifted MDP on the space of probability measures, and establish the dynamic programming Bellman fixed point equation satisfied by the value function, as well as the existence of  $\epsilon$ -optimal randomized feedback controls. The arguments of proof involve an original measurable optimal coupling for the Wasserstein distance. This provides a procedure for learning strategies in a large population of interacting collaborative agents.

**MSC Classification:** 90C40, 49L20.

**Key words:** Mean-field, Markov decision process, conditional propagation of chaos, measurable coupling, randomized control.

---

\*LPSM, Université de Paris medericmotte at gmail.com The author acknowledges support of the DIM MathInnov.

†LPSM, Université de Paris, and CREST-ENSAE, pham at lpsm.paris This work was partially supported by the Chair Finance & Sustainable Development / the FiME Lab (Institut Europlace de Finance)

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>The <math>N</math>-agent and the limiting McKean-Vlasov MDP</b>	<b>5</b>
<b>3</b>	<b>Lifted MDP on <math>\mathcal{P}(\mathcal{X})</math></b>	<b>12</b>
3.1	Case without common noise . . . . .	14
3.2	Case with finite state space $\mathcal{X}$ and with common noise . . . . .	15
<b>4</b>	<b>General case and Bellman fixed point equation in <math>\mathcal{P}(\mathcal{X})</math></b>	<b>18</b>
4.1	A general lifted MDP on $\mathcal{P}(\mathcal{X})$ . . . . .	18
4.2	Bellman fixed point on $\mathcal{P}(\mathcal{X})$ . . . . .	20
4.3	Building $\epsilon$ -optimal randomized feedback controls . . . . .	24
4.4	Relaxing the randomization hypothesis . . . . .	28
4.5	Feedback vs open-loop vs randomized controls . . . . .	30
4.6	Computing value function and $\epsilon$ -optimal strategies in CMKV-MDP . . . . .	32
<b>5</b>	<b>Conclusion</b>	<b>36</b>
<b>A</b>	<b>Some useful results on conditional law</b>	<b>36</b>
<b>B</b>	<b>Wasserstein convergence of the empirical law</b>	<b>37</b>
<b>C</b>	<b>Proof of coupling results</b>	<b>38</b>

## 1 Introduction

Optimal control of McKean-Vlasov (MKV) systems, also known as mean-field control (MFC) problems, has sparked a great interest in the domain of applied probabilities during the last decade. In these optimization problems, the transition dynamics of the system and the reward/gain function depend not only on the state and action of the agent/controller, but also on their probability distributions. These problems are motivated from models of large population of interacting cooperative agents obeying to a social planner (center of decision), and are often justified heuristically as the asymptotic regime with infinite number of agents under Pareto efficiency. Such problems have found numerous applications in distributed energy, herd behavior, finance, etc.

A large literature has already emerged on continuous-time models for the optimal control of McKean-Vlasov dynamics, and dynamic programming principle (in other words time consistency) has been established in this context in the papers [16], [19], [1], [8]. We point out the work [15], which is the first paper to rigorously connect mean-field control to large systems of controlled processes, see also the recent paper [10], and refer to the books [2], [5] for an overview of the subject.

**Our work and main contributions.** In this paper, we introduce a general discrete time framework by providing an exhaustive study of Markov decision process (MDP) under mean-field interaction in the presence of *common noise*, and when optimization is performed over *open-loop controls* on infinite horizon. Such model is called conditional McKean-Vlasov MDP, shortly abbreviated in the sequel as

CMKV-MDP, and the set-up is the mathematical framework for a theory of reinforcement learning with mean-field interaction. Let us first briefly describe and motivate the main features of our framework:

- (i) The controls are open-loop: it emulates the generally unobservable nature of the states in mean-field situations (political opinions, targeted advertising, health conditions). Yet, the noises driving these states (people’s characteristics, habits, surroundings, and navigation data) are nowadays observable.
- (ii) The controls are *individualized*: the action applied to an individual only depends upon her own noise. Indeed, for the strategy to be implementable in practice, the action must be computed by a program running on the individual’s own device (computer, phone), and, to limit data transfers, we impose that it only uses this device’s data.
- (iii) The dynamics of individuals depend upon a common noise, emulating the fact that they are influenced by common information (public data) which may vary over time. By considering an i.i.d. common noise sequence, we shall see that the value function is characterized as the unique fixed point of a Bellman operator, which is the basis for designing algorithms in reinforcement learning.

Compared to continuous-time models, discrete-time McKean-Vlasov control problems have been less studied in the literature. In [18], the authors consider a finite-horizon problem without common noise and state the dynamic programming (Bellman) equation for MFC with closed-loop (also called feedback) controls, that are restricted to depend on the state. Very recently, the works [6], [12] addressed Bellman equations for MFC problems in the context of reinforcement learning. The paper [12] considers relaxed controls in their MFC formulation but without common noise, and derives the Bellman equation for the  $Q$ -value function as a deterministic control problem that we obtain here as a particular case (see our Remark 4.11). The framework in [6] is closest to ours by considering also common noise, however with the following differences: these authors restrict their attention to stationary feedback policies, and reformulate their MFC control problem as a MDP on the space of probability measures by deriving formally (leaving aside the measurability issues and assuming the existence of a stationary feedback control) the associated Bellman equation, which is then used for the development of  $Q$ -learning algorithms. Notice that [6], [12] do not consider dependence upon the probability distribution of the control in the state transition dynamics and reward function.

Besides the introduction of a general framework including a mean-field dependence on the pair state/action, our first main contribution is to rigorously connect CMKV-MDP to a large but finite system of MDP with interacting processes. We prove the almost sure and  $L^1$  conditional propagation of chaos, i.e., the convergence, as the number of interacting agents  $N$  tends to infinity, of the state processes and gains of the  $N$ -individual population control problem towards the corresponding object in the CMKV-MDP. Furthermore, by relying on rate of convergence in Wasserstein distance of the empirical measure, we give a rate of convergence for the limiting CMKV-MDP under suitable Lipschitz assumptions on the state transition and reward functions, which is new to the best of our knowledge.

Our second contribution is to obtain the correspondence of our CMKV-MDP with a suitable lifted MDP on the space of probability measures. Starting from open-loop controls, this is achieved in general by introducing relaxed (i.e. measure-valued) controls in the enlarged state/action space, and by emphasizing the measurability issues arising in the presence of common noise and with continuous state space. In the special case without common noise or with finite state space, the relaxed control in

the lifted MDP is reduced to the usual notion in control theory, also known as mixed or randomized strategies in game theory. While it is known in standard MDP that an optimal control (when it exists) is in pure form, relaxed control appears naturally in MFC where the social planner has to sample the distribution of actions instead of simply assigning the same pure strategy among the population in order to perform the best possible collective gain.

The reformulation of the original problem as a lifted MDP leads us to consider an associated dynamic programming equation written in terms of a Bellman fixed point equation in the space of probability measures. Our third contribution is to establish rigorously the Bellman equation satisfied by the state value function of the CMKV-MDP, and then by the state-action value function, called  $Q$ -function in the reinforcement learning terminology. This is obtained under the crucial assumption that the initial information filtration is generated by an atomless random variable, i.e., that it is rich enough, and calls upon original measurable optimal coupling results for the Wasserstein distance. Moreover, and this is our fourth contribution, the methodology of proof allows us to obtain as a by-product the existence of an  $\epsilon$ -optimal control, which is constructed from randomized feedback policies under a randomization hypothesis. This shows in particular that the value function of CMKV-MDP over open-loop controls is equal to the value function over randomized feedback controls, and we highlight that it may be strictly larger than the value function of CMKV-MDP over “pure” feedback controls, i.e., without randomization. This is a notable difference with respect to the classical (without mean-field dependence) theory of MDP as studied e.g. in [3], [21]. We discuss and illustrate with a set of simple examples the difference of control strategies (open loop vs feedback), and the crucial role of the randomization hypothesis.

Finally, we discuss how to compute the value function and approximate optimal randomized feedback controls from the Bellman equation according to value or policy iteration methods and by discretization of the state space and of the space of probability measures.

**Outline of the paper.** The rest of the paper is organized as follows. Section 2 carefully formulates both the  $N$ -individual model and the CMKV-MDP, and show their connection by providing the rate of convergence of the latter to the limiting MFC when  $N$  goes to infinity. In Section 3, we establish the correspondence of the CMKV-MDP with a lifted MDP on the space of probability measures with usual relaxed controls when there is no common noise or when the state space is finite. In the general case considered in Section 4, we show how to lift the CMKV-MDP by a suitable enlargement of the action space in order to get the correspondance with a MDP on the Wasserstein space. We then derive the associated Bellman fixed point equation satisfied by the value function, and obtain the existence of approximate randomized feedback controls. We also highlight the differences between open-loop vs feedback vs randomized controls. We conclude in Section 5 by indicating some questions for future research. Finally, we collect in the Appendix some useful and technical results including measurable coupling arguments used in the proofs of the paper.

**Notations.** Given two measurable spaces  $(\mathcal{X}_1, \Sigma_1)$  and  $(\mathcal{X}_2, \Sigma_2)$ , we denote by  $\text{pr}_1$  (resp.  $\text{pr}_2$ ) the projection function  $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \mapsto x_1 \in \mathcal{X}_1$  (resp.  $x_2 \in \mathcal{X}_2$ ). For a measurable function  $\Phi : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ , and a positive measure  $\mu_1$  on  $(\mathcal{X}_1, \Sigma_1)$ , the pushforward measure  $\Phi \star \mu_1$  is the measure on  $(\mathcal{X}_2, \Sigma_2)$  defined by

$$\Phi \star \mu_1(B_2) = \mu_1(\Phi^{-1}(B_2)), \quad \forall B_2 \in \Sigma_2.$$

We denote by  $\mathcal{P}(\mathcal{X}_1)$  the set of probability measures on  $\mathcal{X}_1$ , and  $\mathcal{C}(\mathcal{X}_1)$  the cylinder (or weak)  $\sigma$ -algebra on  $\mathcal{P}(\mathcal{X}_1)$ , that is the smallest  $\sigma$ -algebra making all the functions  $\mu \in \mathcal{P}(\mathcal{X}_1) \mapsto \mu(B_1) \in [0, 1]$ ,

measurable for all  $B_1 \in \Sigma_1$ .

A probability kernel  $\nu$  on  $\mathcal{X}_1 \times \mathcal{X}_2$ , denoted  $\nu \in \hat{\mathcal{X}}_2(\mathcal{X}_1)$ , is a measurable mapping from  $(\mathcal{X}_1, \Sigma_1)$  into  $(\mathcal{P}(\mathcal{X}_2), \mathcal{C}(\mathcal{X}_2))$ , and we shall write indifferently  $\nu(x_1, B_2) = \nu(x_1)(B_2)$ , for all  $x_1 \in \mathcal{X}_1$ ,  $B_2 \in \Sigma_2$ . Given a probability measure  $\mu_1$  on  $(\mathcal{X}_1, \Sigma_1)$ , and a probability kernel  $\nu \in \hat{\mathcal{X}}_2(\mathcal{X}_1)$ , we denote by  $\mu_1 \cdot \nu$  the probability measure on  $(\mathcal{X}_1 \times \mathcal{X}_2, \Sigma_1 \otimes \Sigma_2)$  defined by

$$(\mu_1 \cdot \nu)(B_1 \times B_2) = \int_{B_1 \times B_2} \mu_1(dx_1) \nu(x_1, dx_2), \quad \forall B_1 \in \Sigma_1, B_2 \in \Sigma_2.$$

Let  $X_1$  and  $X_2$  be two random variables valued respectively on  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , denoted  $X_i \in L^0(\Omega; \mathcal{X}_i)$ . We denote by  $\mathcal{L}(X_i)$  the probability distribution of  $X_i$ , and by  $\mathcal{L}(X_2|X_1)$  the conditional probability distribution of  $X_2$  given  $X_1$ . With these notations, when  $X_2 = \Phi(X_1)$ , then  $\mathcal{L}(X_2) = \Phi \star \mathcal{L}(X_1)$ .

When  $(\mathcal{Y}, d)$  is a compact metric space, the set  $\mathcal{P}(\mathcal{Y})$  of probability measures on  $\mathcal{Y}$  is equipped with the Wasserstein distance

$$\mathcal{W}(\mu, \mu') = \inf \left\{ \int_{\mathcal{Y}^2} d(y, y') \boldsymbol{\mu}(dy, dy') : \boldsymbol{\mu} \in \boldsymbol{\Pi}(\mu, \mu') \right\},$$

where  $\boldsymbol{\Pi}(\mu, \mu')$  is the set of probability measures on  $\mathcal{Y} \times \mathcal{Y}$  with marginals  $\mu$  and  $\mu'$ , i.e.,  $\text{pr}_1 \star \boldsymbol{\mu} = \mu$ , and  $\text{pr}_2 \star \boldsymbol{\mu} = \mu'$ . Since  $(\mathcal{Y}, d)$  is compact, it is known (see e.g. Corollary 6.13 in [23]) that the Borel  $\sigma$ -algebra generated by the Wasserstein metric coincides with the cylinder  $\sigma$ -algebra on  $\mathcal{P}(\mathcal{Y})$ , i.e., Wasserstein distances metrize weak convergence. We also recall the dual Kantorovich-Rubinstein representation of the Wasserstein distance

$$\mathcal{W}(\mu, \mu') = \sup \left\{ \int_{\mathcal{Y}} \phi d(\mu - \mu') : \phi \in Lip(\mathcal{Y}; \mathbb{R}), [\phi]_{lip} \leq 1 \right\},$$

where  $Lip(\mathcal{Y}; \mathbb{R})$  is the set of Lipschitz continuous functions  $\phi$  from  $\mathcal{Y}$  into  $\mathbb{R}$ , and  $[\phi]_{lip} = \sup\{|\phi(y) - \phi(y')|/d(y, y') : y, y' \in \mathcal{Y}, y \neq y'\}$ .

## 2 The $N$ -agent and the limiting McKean-Vlasov MDP

We formulate the mean-field Markov Decision Process (MDP) in a large population model with indistinguishable agents  $i \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$ .

Let  $\mathcal{X}$  (the state space) and  $A$  (the action space) be two compact Polish spaces equipped respectively with their metric  $d$  and  $d_A$ . We denote by  $\mathcal{P}(\mathcal{X})$  (resp.  $\mathcal{P}(A)$ ) the space of probability measures on  $\mathcal{X}$  (resp.  $A$ ) equipped respectively with their Wasserstein distance  $\mathcal{W}$  and  $\mathcal{W}_A$ . We also consider the product space  $\mathcal{X} \times A$ , equipped with the metric  $\mathbf{d}((x, a), (x', a')) = d(x, x') + d_A(a, a')$ ,  $x, x' \in \mathcal{X}$ ,  $a, a' \in A$ , and the associated space of probability measure  $\mathcal{P}(\mathcal{X} \times A)$ , equipped with its Wasserstein distance  $\mathbf{W}$ . Let  $G$ ,  $E$ , and  $E^0$  be three measurable spaces, representing respectively the initial information, idiosyncratic noise, and common noise spaces.

We denote by  $\Pi_{OL}$  the set of sequences  $(\pi_t)_{t \in \mathbb{N}}$  (called *open-loop policies*) where  $\pi_t$  is a measurable function from  $G \times E^t \times (E^0)^t$  into  $A$  for  $t \in \mathbb{N}$ .

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space on which are defined the following family of mutually i.i.d. random variables

- $(\Gamma^i, \xi^i)_{i \in \mathbb{N}^*}$  (initial informations and initial states) valued in  $G \times \mathcal{X}$
- $(\varepsilon_t^i)_{i \in \mathbb{N}^*, t \in \mathbb{N}}$  (idiosyncratic noises) valued in  $E$  with probability distribution  $\lambda_\varepsilon$

- $\varepsilon^0 := (\varepsilon_t^0)_{t \in \mathbb{N}}$  (common noise) valued in  $E^0$ .

We assume that  $\mathcal{F}$  contains an atomless random variable, i.e.,  $\mathcal{F}$  is rich enough, so that any probability measure  $\nu$  on  $\mathcal{X}$  (resp.  $A$  or  $\mathcal{X} \times A$ ) can be represented by the law of some random variable  $Y$  on  $\mathcal{X}$  (resp.  $A$  or  $\mathcal{X} \times A$ ), and we write  $Y \sim \nu$ , i.e.,  $\mathcal{L}(Y) = \nu$ . Given an open-loop policy  $\pi$ , we associate an *open-loop control* for individual  $i \in \mathbb{N}^*$  as the process  $\alpha^{i,\pi}$  defined by

$$\alpha^{i,\pi} = \pi_t(\Gamma^i, (\varepsilon_s^i)_{s \leq t}, (\varepsilon_s^0)_{s \leq t}), \quad t \in \mathbb{N}.$$

In other words, an open-loop control is a non-anticipative process that depends on the initial information, the past idiosyncratic and common noises, but not on the states of the agent in contrast with closed-loop control.

Given  $N \in \mathbb{N}^*$ , and  $\pi \in \Pi_{OL}$ , the state process of agent  $i = 1, \dots, N$  in an  $N$ -agent MDP is given by the dynamical system

$$\begin{cases} X_0^{i,N,\pi} &= \xi^i \\ X_{t+1}^{i,N,\pi} &= F(X_t^{i,N,\pi}, \alpha_t^{i,\pi}, \frac{1}{N} \sum_{j=1}^N \delta_{(X_t^{j,N,\pi}, \alpha_t^{j,\pi})}, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \end{cases}$$

where  $F$  is a measurable function from  $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E \times E^0$  into  $\mathcal{X}$ , called state transition function. The  $i$ -th individual contribution to the influencer's gain over an infinite horizon is defined by

$$J_i^{N,\pi} := \sum_{t=0}^{\infty} \beta^t f\left(X_t^{i,N,\pi}, \alpha_t^{i,\pi}, \frac{1}{N} \sum_{j=1}^N \delta_{(X_t^{j,N,\pi}, \alpha_t^{j,\pi})}\right), \quad i = 1, \dots, N,$$

where the reward  $f$  is a measurable real-valued function on  $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A)$ , assumed to be bounded (recall that  $\mathcal{X}$  and  $A$  are compact spaces), and  $\beta$  is a positive discount factor in  $[0, 1)$ . The influencer's renormalized and expected gains are

$$J^{N,\pi} := \frac{1}{N} \sum_{i=1}^N J_i^{N,\pi}, \quad V^{N,\pi} := \mathbb{E}[J^{N,\pi}],$$

and the optimal value of the influencer is  $V^N := \sup_{\pi \in \Pi_{OL}} V^{N,\pi}$ . Observe that the agents are indistinguishable in the sense that the initial pair of information/state  $(\Gamma^i, \xi^i)_i$ , and idiosyncratic noises are i.i.d., and the state transition function  $F$ , reward function  $f$ , and discount factor  $\beta$  do not depend on  $i$ .

Let us now consider the asymptotic problem when the number of agents  $N$  goes to infinity. In view of the propagation of chaos argument, we expect that the state process of agent  $i \in \mathbb{N}^*$  in the infinite population model to be governed by the conditional McKean-Vlasov dynamics

$$\begin{cases} X_0^{i,\pi} &= \xi^i \\ X_{t+1}^{i,\pi} &= F(X_t^{i,\pi}, \alpha_t^{i,\pi}, \mathbb{P}_{(X_t^{i,\pi}, \alpha_t^{i,\pi})}^0, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}. \end{cases} \quad (2.1)$$

Here, we denote by  $\mathbb{P}^0$  and  $\mathbb{E}^0$  the conditional probability and expectation knowing the common noise  $\varepsilon^0$ , and then, given a random variable  $Y$  valued in  $\mathcal{Y}$ , we denote by  $\mathbb{P}_Y^0$  or  $\mathcal{L}^0(Y)$  its conditional law knowing  $\varepsilon^0$ , which is a random variable valued in  $\mathcal{P}(\mathcal{Y})$  (see Lemma A.2). The  $i$ -th individual contribution to the influencer's gain in the infinite population model is

$$J_i^\pi := \sum_{t=0}^{\infty} \beta^t f\left(X_t^{i,\pi}, \alpha_t^{i,\pi}, \mathbb{P}_{(X_t^{i,\pi}, \alpha_t^{i,\pi})}^0\right), \quad i \in \mathbb{N}^*,$$

and we define the conditional gain, expected gain, and optimal value, respectively by

$$\begin{aligned} J^\pi &:= \mathbb{E}^0[J_i^\pi] = \mathbb{E}^0[J_1^\pi], \quad i \in \mathbb{N}^* \quad (\text{by indistinguishability of the agents}), \\ V^\pi &:= \mathbb{E}[J^\pi], \quad V := \sup_{\pi \in \Pi_{OL}} V^\pi. \end{aligned} \quad (2.2)$$

Problem (2)-(2) is called conditional McKean-Vlasov Markov decision process, CMKV-MDP in short.

The main goal of this Section is to rigorously connect the finite-agent model to the infinite population model with mean-field interaction by proving the convergence of the  $N$ -agent MDP to the CMKV-MDP. First, we prove the almost sure convergence of the state processes under some continuity assumptions on the state transition function.

**Proposition 2.1** *Assume that for all  $(x_0, a, \nu_0, e^0) \in \mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E^0$ , the (random) function*

$$(x, \nu) \in \mathcal{X} \times \mathcal{P}(\mathcal{X} \times A) \mapsto F(x, a, \nu, \varepsilon_1^1, e^0) \in \mathcal{X} \quad \text{is continuous in } (x_0, \nu_0) \text{ a.s.}$$

*i.e.  $\mathbb{P}\left[\lim_{(x, \nu) \rightarrow (x_0, \nu_0)} F(x, a, \nu, \varepsilon_1^1, e^0) = F(x_0, a, \nu_0, \varepsilon_1^1, e^0)\right] = 1$ . Then, for any  $\pi \in \Pi_{OL}$ ,  $X_t^{i, N, \pi} \xrightarrow[N \rightarrow \infty]{a.s.} X_t^{i, \pi}$ ,  $i \in \mathbb{N}^*$ ,  $t \in \mathbb{N}$ , and*

$$\mathbf{W}\left(\frac{1}{N} \sum_{i=1}^N \delta_{(X_t^{i, N, \pi}, \alpha_t^{i, \pi})}, \mathbb{P}_{(X_t^{i, \pi}, \alpha_t^{i, \pi})}^0\right) \xrightarrow[N \rightarrow \infty]{a.s.} 0, \quad \frac{1}{N} \sum_{i=1}^N d(X_t^{i, N, \pi}, X_t^{i, \pi}) \xrightarrow[N \rightarrow \infty]{} 0 \quad a.s.$$

*Furthermore, if we assume that for all  $a \in A$ , the function  $(x, \nu) \in \mathcal{X} \times \mathcal{P}(\mathcal{X} \times A) \mapsto f(x, a, \nu)$  is continuous, then*

$$J_i^{N, \pi} \xrightarrow[N \rightarrow \infty]{a.s.} J_i^\pi, \quad J^{N, \pi} \xrightarrow[N \rightarrow \infty]{a.s.} J^\pi, \quad V^{N, \pi} \xrightarrow[N \rightarrow \infty]{} V^\pi, \quad \text{and} \quad \liminf_{N \rightarrow \infty} V^N \geq V.$$

**Remark 2.1** We stress the importance of making the continuity assumption for  $F$  only a.s. w.r.t. the noise  $\varepsilon_1^1$ . Indeed, most of the practical applications we have in mind concerns finite spaces  $\mathcal{X}$ , for which  $F$  cannot be, strictly speaking, continuous, since it maps from a continuous space  $\mathcal{P}(\mathcal{X} \times A)$  to a finite space  $\mathcal{X}$ . However, in this case, the discontinuities will essentially be placed on the boundaries between each region where  $F$  is constant. Our assumption then simply formalizes the idea that these boundaries depend upon  $\varepsilon_1^1$  in a way that no point  $(x_0, a, \nu_0, e^0) \in \mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E^0$  has a positive probability (w.r.t.  $\varepsilon_1^1$ ) to be on a boundary. In other words, we assume that the discontinuity sets always vary with  $\varepsilon_1^1$ , which is a natural phenomenon.  $\square$

**Proof.** Fix  $\pi \in \Pi_{OL}$ . We omit the dependence of the state processes and control on  $\pi$ , and denote by  $\nu_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^{i, N}, \alpha_t^i)}$ ,  $\nu_t^{N, \infty} := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^i, \alpha_t^i)}$ , and  $\nu_t := \mathbb{P}_{(X_t^i, \alpha_t^i)}^0$ .

(1) We first prove the convergence of trajectories, for all  $i \in \mathbb{N}^*$ ,

$$\mathbb{P}\left[\lim_{N \rightarrow \infty} (X_t^{i, N}, \nu_t^N) = (X_t^i, \nu_t)\right] = 1, \quad \mathbb{P}\left[\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N d(X_t^{j, N}, X_t^j) = 0\right] = 1,$$

by induction on  $t \in \mathbb{N}$ .

- *Initialization.* For  $t = 0$ , we have  $X_0^{i, N} = \xi^i = X_0^i$ ,  $\alpha_0^i = \pi_0(\Gamma_i)$ , for all  $N \in \mathbb{N}_*$  and  $i \in \mathbb{N}^*$ , which obviously implies that  $X_0^{i, N} \xrightarrow[N \rightarrow \infty]{a.s.} X_0^i$  and  $\frac{1}{N} \sum_{j=1}^N d(X_0^{j, N}, X_0^j) \xrightarrow[N \rightarrow \infty]{} 0$ . Moreover,  $\mathcal{W}(\nu_0^N, \nu_0)$



=  $\mathcal{W}(\nu_0^{N,\infty}, \nu_0)$ , which converges to zero a.s., when  $N$  goes to infinity, by the weak convergence of empirical measures (see [22]), and the fact that Wasserstein distance metrizes weak convergence.

- *Hereditary Property.* We have

$$X_{t+1}^{i,N} = F(X_t^{i,N}, \alpha_t^i, \nu_t^N, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0) \quad \text{and} \quad X_{t+1}^i = F(X_t^i, \alpha_t^i, \nu_t, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0). \quad (2.3)$$

By a simple conditioning, we notice that  $\mathbb{P}[\lim_{N \rightarrow \infty} X_{t+1}^{i,N} = X_{t+1}^i] = \mathbb{E}[P((X_t^{i,N}, \nu_t^N)_N, X_t^i, \nu_t, \alpha_t^i)]$ , where

$$P((x_N, \nu_N)_N, x, \nu, a) = \mathbb{P}\left[\lim_{N \rightarrow \infty} F(x_N, a, \nu_N, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0) = F(x, a, \nu, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0)\right].$$

By the continuity assumption on  $F$ , we see that  $P((x_N, \nu_N)_N, x, \nu, a)$  is bounded from below by  $\mathbf{1}_{\lim_{N \rightarrow \infty} (x_N, \nu_N) = (x, \nu)}$ , and thus

$$\mathbb{P}\left[\lim_{N \rightarrow \infty} X_{t+1}^{i,N} = X_{t+1}^i\right] \geq \mathbb{P}\left[\lim_{N \rightarrow \infty} (X_t^{i,N}, \nu_t^N) = (X_t^i, \nu_t)\right].$$

This proves by induction hypothesis that  $\mathbb{P}\left[\lim_{N \rightarrow \infty} X_{t+1}^{i,N} = X_{t+1}^i\right] = 1$ .

Let us now show that  $\frac{1}{N} \sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i) \xrightarrow[N \rightarrow \infty]{a.s.} 0$ . From (2), we have

$$d(X_{t+1}^{i,N}, X_{t+1}^i) \leq \Delta_{D_N} F(X_t^i, \alpha_t^i, \nu_t, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0)$$

with  $D_N := \max[d(X_t^{i,N}, X_t^i), \mathbf{W}(\nu_t^N, \nu_t)]$ , and

$$\Delta_y F(x, a, \nu, e, e^0) := \sup_{(x', \nu') \in D} \{d(F(x', a, \nu', e, e^0), F(x, a, \nu, e, e^0)) \mathbf{1}_{\max[d(x', x), \mathbf{W}(\nu', \nu)] \leq y}\},$$

where  $D$  is a fixed countable dense set of the separable space  $\mathcal{X} \times \mathcal{P}(\mathcal{X} \times A)$ , which implies that  $(y, x, a, \nu, e, e^0) \mapsto \Delta_y F(x, a, \nu, e, e^0)$  is a measurable function. Fix  $\epsilon > 0$ . Let  $\Delta_{\mathcal{X}}$  denote the diameter of the compact metric space  $\mathcal{X}$ . We thus have, for any  $\eta > 0$ ,

$$\begin{aligned} d(X_{t+1}^{i,N}, X_{t+1}^i) &\leq d(X_{t+1}^{i,N}, X_{t+1}^i) \mathbf{1}_{D_N \leq \eta} + d(X_{t+1}^{i,N}, X_{t+1}^i) \mathbf{1}_{D_N > \eta} \\ &\leq \Delta_{\eta} F(X_t^i, \alpha_t^i, \nu_t, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0) + \Delta_{\mathcal{X}} \mathbf{1}_{d(X_t^{i,N}, X_t^i) > \eta} + \Delta_{\mathcal{X}} \mathbf{1}_{\mathbf{W}(\nu_t^N, \nu_t) > \eta}, \end{aligned}$$

and thus

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i) \\ &\leq \frac{1}{N} \sum_{i=1}^N \Delta_{\eta} F(X_t^i, \alpha_t^i, \nu_t, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0) + \frac{\Delta_{\mathcal{X}}}{\eta N} \sum_{i=1}^N d(X_t^{i,N}, X_t^i) + \Delta_{\mathcal{X}} \mathbf{1}_{\mathbf{W}(\nu_t^N, \nu_t) > \eta}. \end{aligned}$$

The second and third terms in the right hand side converge to 0 by induction hypothesis, and by Proposition B.1, the first term converges to

$$\mathbb{E}^0[\Delta_{\eta} F(X_t^1, \alpha_t^1, \nu_t, \varepsilon_{t+1}^1, \varepsilon_{t+1}^0)] = \mathbb{E}^0\left[\mathbb{E}\left[\Delta_{\eta} F(x_t^1, a_t^1, \nu_t, \varepsilon_{t+1}^1, e_{t+1}^0)\right]_{(x_t^1, a_t^1, e_{t+1}^0) := (X_t^1, \alpha_t^1, \varepsilon_{t+1}^0)}\right].$$

As  $\eta \rightarrow 0$ , the inner expectation tends to zero by continuity assumption on  $F$  and by dominated convergence. Then, the outer expectation converges to zero by conditional dominated convergence,

and will thus be smaller than  $\frac{\epsilon}{2}$  for  $\eta$  small enough, which implies that  $\frac{1}{N} \sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i)$  will be smaller than  $\epsilon$  for  $N$  large enough.

Let us finally prove that  $\mathbf{W}(\nu_{t+1}^N, \nu_{t+1}) \xrightarrow[N \rightarrow \infty]{a.s.} 0$ . We have  $\mathbf{W}(\nu_{t+1}^N, \nu_{t+1}) \leq \mathbf{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty}) + \mathbf{W}(\nu_{t+1}^{N,\infty}, \nu_{t+1})$ . To dominate the first term  $\mathbf{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty})$ , notice that, given a variable  $U_N \sim \mathcal{U}(\{1, \dots, N\})$ , the random measure  $\nu_{t+1}^N$  (resp.  $\nu_{t+1}^{N,\infty}$ ) is, at fixed  $\omega \in \Omega$ , the law of the pair of random variable  $(X_{t+1}^{U_N, N}(\omega), \alpha_{t+1}^{U_N}(\omega))$  (resp.  $(X_{t+1}^{U_N}(\omega), \alpha_{t+1}^{U_N}(\omega))$ ) where we stress that only  $U_N$  is random here, essentially selecting each sample of these empirical measures with probability  $\frac{1}{N}$ . Thus, by definition of the Wasserstein distance,  $\mathbf{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty})$  is dominated by  $\mathbb{E}[d(X_{t+1}^{U_N, N}(\omega), X_{t+1}^{U_N}(\omega))] = \frac{1}{N} \sum_{i=1}^N d(X_{t+1}^{i,N}(\omega), X_{t+1}^i(\omega))$ , which has been proved to converge to zero. For the second term, observe that  $\alpha_{t+1}^i = \pi_{t+1}(\Gamma^i, (\varepsilon_s^i)_{s \leq t+1}, (\varepsilon_s^0)_{s \leq t+1})$ , and by Proposition A.1, there exists a measurable function  $f_{t+1} : \mathcal{X} \times G \times E^{t+1} \times (E^0)^{t+1} \rightarrow \mathcal{X}$  such that  $X_{t+1}^{i,N} = f_{t+1}(\Gamma^i, (\varepsilon_s^i)_{s \leq t+1}, (\varepsilon_s^0)_{s \leq t+1})$ . From Proposition B.1, we then deduce that  $\mathbf{W}(\nu_{t+1}^{N,\infty}, \nu_{t+1})$  converges to zero as  $N$  goes to infinity. This concludes the induction.

(2) Let us now study the convergence of gains. By the continuity assumption on  $f$ , we have  $f(X_t^{i,N}, \alpha_t^i, \nu_t^N) \xrightarrow[N \rightarrow \infty]{a.s.} f(X_t^i, \alpha_t^i, \nu_t)$  for all  $t \in \mathbb{N}$ . Thus, as  $f$  is bounded, we get by dominated convergence that  $J_i^{N,\pi} \xrightarrow[N \rightarrow \infty]{a.s.} J_i^\pi$ .

Let us now study the convergence of  $J^{N,\pi}$  to  $J^\pi$ . We write

$$|J^{N,\pi} - J^\pi| \leq \frac{1}{N} \sum_{i=1}^N |J_i^{N,\pi} - J_i^\pi| + \left| \frac{1}{N} \sum_{i=1}^N J_i^\pi - J^\pi \right| =: S_N^1 + S_N^2.$$

The second term  $S_N^2$  converges a.s. to zero by Propositions A.1 and B.1, as  $N$  goes to infinity. On the other hand,

$$S_N^1 \leq \sum_{t=0}^{\infty} \beta^t \Delta_N(f), \quad \text{with } \Delta_N(f) := \frac{1}{N} \sum_{i=1}^N |f(X_t^{i,N}, \alpha_t^i, \nu_t^N) - f(X_t^i, \alpha_t^i, \nu_t)|.$$

By the same argument as above in (1) for showing that  $\frac{1}{N} \sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i) \xrightarrow[N \rightarrow \infty]{a.s.} 0$ , we prove that  $\Delta_N(f)$  tends a.s. to zero as  $N \rightarrow \infty$ . Since  $f$  is bounded, we deduce by the dominated convergence theorem that  $S_N^1$  converges a.s. to zero as  $N$  goes to infinity, and thus  $J^{N,\pi} \xrightarrow[N \rightarrow \infty]{a.s.} J^\pi$ . By dominated convergence, we then also obtain that  $V^{N,\pi} = \mathbb{E}[J^N] \xrightarrow[N \rightarrow \infty]{a.s.} \mathbb{E}[J^\pi] = V^\pi$ . Finally, by considering an  $\epsilon$ -optimal policy  $\pi_\epsilon$  for  $V$ , we have

$$\liminf_{N \rightarrow \infty} V^N \geq \lim_{N \rightarrow \infty} V^{N,\pi_\epsilon} = V^{\pi_\epsilon} \geq V - \epsilon,$$

which implies, by sending  $\epsilon$  to zero, that  $\liminf_{N \rightarrow \infty} V^N \geq V$ .  $\square$

Next, under Lipschitz assumptions on the state transition and reward functions, we prove the corresponding convergence in  $\mathbb{L}^1$ , which implies the convergence of the optimal value, and also a rate of convergence in terms of the rate of convergence in Wasserstein distance of the empirical measure.

**(HF<sub>lip</sub>)** There exists  $K_F > 0$ , such that for all  $a \in A$ ,  $e^0 \in E^0$ ,  $x, x' \in \mathcal{X}$ ,  $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$ ,

$$\mathbb{E}[d(F(x, a, \nu, \varepsilon_1^1, e^0), F(x', a, \nu', \varepsilon_1^1, e^0))] \leq K_F(d(x, x') + \mathbf{W}(\nu, \nu')).$$

**(Hf<sub>lip</sub>)** There exists  $K_f > 0$ , such that for all  $a \in A$ ,  $x, x' \in \mathcal{X}$ ,  $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$ ,

$$d(f(x, a, \nu), f(x', a, \nu')) \leq K_f(d(x, x') + \mathbf{W}(\nu, \nu')).$$

**Remark 2.2** Here again, we stress the importance of making the regularity assumptions for  $F$  in *expectation* only. For the same argument as in Remark 2.1, when  $\mathcal{X}$  is finite,  $F$  cannot be, strictly speaking, Lipschitz. However,  $F$  can be Lipschitz *in expectation*, e.g. once integrated w.r.t. the idiosyncratic noise, and, again, it is a very natural assumption.  $\square$

In the sequel, we shall denote by  $\Delta_{\mathcal{X}}$  the diameter of the metric space  $\mathcal{X}$ , and define

$$M_N := \sup_{\nu \in \mathcal{P}(\mathcal{X} \times A)} \mathbb{E}[\mathbf{W}(\nu_N, \nu)], \quad (2.9)$$

where  $\nu_N$  is the empirical measure  $\nu_N = \frac{1}{N} \sum_{n=1}^N \delta_{Y_n}$ ,  $(Y_n)_{1 \leq n \leq N}$  are i.i.d. random variables with law  $\nu$ . We recall in the next Lemma recent results about non asymptotic bounds for the mean rate of convergence in Wasserstein distance of the empirical measure.

**Lemma 2.1** *We have  $M_N \xrightarrow{N \rightarrow \infty} 0$ . Furthermore,*

- *If  $\mathcal{X} \times A \subset \mathbb{R}^d$  for some  $d \in \mathbb{N}^*$ , then:  $M_N = \mathcal{O}(N^{-\frac{1}{2}})$  for  $d = 1$ ,  $M_N = \mathcal{O}(N^{-\frac{1}{2}} \log(1 + N))$  for  $d = 2$ , and  $M_N = \mathcal{O}(N^{-\frac{1}{d}})$  for  $d \geq 3$ .*
- *If for all  $\delta > 0$ , the smallest number of balls with radius  $\delta$  covering the compact metric set  $\mathcal{X} \times A$  with diameter  $\Delta_{\mathcal{X} \times A}$  is smaller than  $\mathcal{O}\left(\left(\frac{\Delta_{\mathcal{X} \times A}}{\delta}\right)^\theta\right)$  for  $\theta > 2$ , then  $M_N = \mathcal{O}(N^{-1/\theta})$ .*

**Proof.** The second point is proved in [11], and the third one in [4].  $\square$

**Remark 2.3** In the case where the state and action spaces  $\mathcal{X}$  and  $A$  are finite, let  $\phi : \mathcal{X} \times A \hookrightarrow \mathbb{R}$  be any injective function. Then  $\phi^{-1}$  is necessarily Lipschitz. From the dual Kantorovich representation of the Wasserstein distance, we have

$$\begin{aligned} M_N &\leq \sup_{\nu \in \mathcal{P}(\mathcal{X} \times A)} \mathbb{E} \left[ \sup_{g \in L_{\text{lip}}(\mathcal{X} \times A; \mathbb{R}), [g]_{\text{lip}} \leq 1} \int_{\mathcal{X} \times A} g d(\nu_N - \nu) \right] \\ &\leq \sup_{\nu \in \mathcal{P}(\mathcal{X} \times A)} \mathbb{E} \left[ \sup_{g \in L_{\text{lip}}(\mathcal{X} \times A; \mathbb{R}), [g]_{\text{lip}} \leq 1} \int_{\mathcal{X} \times A} g \circ \phi^{-1} d((\phi \star \nu)_N - \phi \star \nu) \right] \\ &\leq C \sup_{\nu \in \mathcal{P}(\mathcal{X} \times A)} \mathbb{E} [\mathcal{W}((\phi \star \nu)_N, \phi \star \nu)] \leq C \sup_{\nu \in \mathcal{P}(\phi(\mathcal{X} \times A))} \mathbb{E} [\mathcal{W}(\nu_N, \nu)], \end{aligned}$$

where  $C$  is the Lipschitz constant of  $\phi^{-1}$ . Thus, by the second point in Lemma 2.1,  $M_N = \mathcal{O}(N^{-\frac{1}{2}})$ .  $\square$

**Theorem 2.1** *Assume  $(\mathbf{HF}_{\text{lip}})$ . For all  $i \in \mathbb{N}^*$ ,  $t \in \mathbb{N}$ ,*

$$\sup_{\pi \in \Pi_{OL}} \mathbb{E} [d(X_t^{i,N,\pi}, X_t^{i,\pi})] = \mathcal{O}(M_N), \quad (2.10)$$

$$\sup_{\pi \in \Pi_{OL}} \mathbb{E} \left[ \mathbf{W} \left( \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^{i,N,\pi}, \alpha^{i,\pi})}, \mathbb{P}^0_{(X_t^{i,\pi}, \alpha^{i,\pi})} \right) \right] = \mathcal{O}(M_N). \quad (2.11)$$

*Furthermore, if we assume  $(\mathbf{Hf}_{\text{lip}})$ , and set  $\gamma = \min \left[ 1, \frac{|\ln \beta|}{(\ln 2K_F)_+} \right]$ , there exists a constant  $C = C(K_F, K_f, \beta, \gamma)$  (explicit in the proof) such that for  $N$  large enough,*

$$\sup_{\pi \in \Pi_{OL}} |V^{N,\pi} - V^\pi| \leq CM_N^\gamma,$$

and thus  $|V^N - V| = \mathcal{O}(M_N^\gamma)$ . Consequently, any  $\varepsilon$ -optimal policy for the CMKV-MDP is  $\mathcal{O}(\varepsilon)$ -optimal for the  $N$ -agent MDP problem, and conversely, any  $\varepsilon$ -optimal policy for the  $N$ -agent MDP problem is  $\mathcal{O}(\varepsilon)$ -optimal for the CMKV-MDP, for  $N$  large enough, namely  $M_N^\gamma = \mathcal{O}(\varepsilon)$ .

**Proof.** Given  $\pi \in \Pi_{OL}$ , denote by  $\nu_t^{N,\pi} := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^{i,N,\pi}, \alpha_t^{i,\pi})}$ ,  $\nu_t^{N,\infty,\pi} := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^{i,\pi}, \alpha_t^{i,\pi})}$  and  $\nu_t^\pi := \mathbb{P}_{(X_t^{i,\pi}, \alpha_t^i)}^0$ . By definition,  $\alpha_t^{i,\pi} = \pi_t(\Gamma^i, (\varepsilon_s^i)_{s \leq t}, (\varepsilon_s^0)_{s \leq t})$ , and by Lemma A.1, we have  $X_t^{i,\pi} = f_t^\pi(\xi^i, \Gamma^i, (\varepsilon_s^i)_{s \leq t}, (\varepsilon_s^0)_{s \leq t})$  for some measurable function  $f_t^\pi \in L^0(\mathcal{X} \times G \times E^t \times (E^0)^t, \mathcal{X})$ . By definition of  $M_N$  in (2), we have

$$\mathbb{E}[\mathbf{W}(\nu_t^{N,\infty,\pi}, \nu_t^\pi)] \leq M_N, \quad \forall N \in \mathbb{N}, \forall \pi \in \Pi_{OL}. \quad (2.13)$$

Let us now prove (2.1) by induction on  $t \in \mathbb{N}$ . At  $t = 0$ ,  $X_0^{i,N,\pi} = X_0^{i,\pi} = \xi^i$ , and the result is obvious. Now assume that it holds true at time  $t \in \mathbb{N}$  and let us show that it then holds true at time  $t + 1$ . By a simple conditioning argument,  $\mathbb{E}[d(X_{t+1}^{i,N,\pi}, X_{t+1}^{i,\pi})] = \mathbb{E}[\Delta(X_t^{i,N,\pi}, X_t^{i,\pi}, \alpha_t^i, \nu_t^{N,\pi}, \nu_t^\pi, \varepsilon_{t+1}^0)]$ , where

$$\begin{aligned} \Delta(x, x', a, \nu, \nu', e^0) &= \mathbb{E}[d(F(x, a, \nu, \varepsilon_{t+1}^i, e^0), F(x', a, \nu', \varepsilon_{t+1}^i, e^0))] \\ &\leq K_F(d(x, x') + \mathbf{W}(\nu, \nu')), \end{aligned} \quad (2.14)$$

by (**HF<sub>lip</sub>**). On the other hand, we have

$$\begin{aligned} \mathbb{E}[\mathbf{W}(\nu_t^{N,\pi}, \nu_t^\pi)] &\leq \mathbb{E}[\mathbf{W}(\nu_t^{N,\pi}, \nu_t^{N,\infty,\pi})] + \mathbb{E}[\mathbf{W}(\nu_t^{N,\infty,\pi}, \nu_t^\pi)] \\ &\leq \mathbb{E}[d(X_t^{i,N,\pi}, X_t^{i,\pi})] + M_N, \end{aligned} \quad (2.15)$$

where we used the fact that  $\mathbf{W}(\nu_t^{N,\pi}, \nu_t^{N,\infty,\pi}) \leq \frac{1}{N} \sum_{i=1}^N d(X_t^{i,N,\pi}, X_t^{i,\pi})$ , and (2). It follows from (2) that

$$\mathbb{E}[d(X_{t+1}^{i,N,\pi}, X_{t+1}^{i,\pi})] \leq K_F \left( 2\mathbb{E}[d(X_t^{i,N,\pi}, X_t^{i,\pi})] + M_N \right), \quad \forall \pi \in \Pi_{OL}, \quad (2.16)$$

which proves that  $\sup_{\pi \in \Pi_{OL}} \mathbb{E}[d(X_{t+1}^{i,N,\pi}, X_{t+1}^{i,\pi})] = \mathcal{O}(M_N)$  by induction hypothesis, and thus (2.1). Plugging (2.1) into (2) then yields (2.1).

Let us now prove the convergence of gains. From (**Hf<sub>lip</sub>**), and (2), we have

$$\begin{aligned} |V^{N,\pi} - V^\pi| &\leq K_f \sum_{t=0}^{\infty} \beta^t \mathbb{E} \left[ d(X_t^{i,N,\pi}, X_t^{i,\pi}) + \mathbf{W}(\nu_t^{N,\pi}, \nu_t^\pi) \right] \\ &\leq K_f \left( 2 \sum_{t=0}^{\infty} \beta^t \delta_t^N + \frac{M_N}{1-\beta} \right), \quad \forall \pi \in \Pi_{OL}, \end{aligned} \quad (2.17)$$

where we set  $\delta_t^N := \sup_{\pi \in \Pi_{OL}} \mathbb{E}[d(X_t^{i,N,\pi}, X_t^{i,\pi})]$ . From (2), we have:  $\delta_{t+1}^N \leq 2K_F \delta_t^N + K_F M_N$ ,  $t \in \mathbb{N}$ , with  $\delta_0^N = 0$ , and so by induction:

$$\delta_t^N \leq \frac{K_F}{1-2K_F} M_N + s_t \left( \frac{K_F}{|2K_F-1|} M_N \right), \quad s_t(m) := m(2K_F)^t, \quad m \geq 0.$$

where we may assume w.l.o.g. that  $2K_F \neq 1$ . Observing that we obviously have  $\delta_t^N \leq \Delta_{\mathcal{X}}$  (the diameter of  $\mathcal{X}$ ), we deduce that

$$\begin{aligned} \sum_{t=0}^{\infty} \beta^t \delta_t^N &\leq \frac{K_F}{(1-2K_F)(1-\beta)} M_N + S \left( \frac{K_F}{|2K_F-1|} M_N \right) \\ S(m) &:= \sum_{t=0}^{\infty} \beta^t \min [s_t(m); \Delta_{\mathcal{X}}], \quad m \geq 0. \end{aligned} \quad (2.18)$$

If  $2\beta K_F < 1$ , we clearly have

$$S(m) \leq \sum_{t=0}^{\infty} (m\beta 2K_F)^t = \frac{m}{1 - \beta 2K_F},$$

and so by (2),  $|V^{N,\pi} - V^\pi| = \mathcal{O}(M_N)$ . Let us now study the case  $2\beta K_F > 1$ . In this case, in particular,  $2K_F > 1$ , thus  $t \mapsto s_t(m)$  is nondecreasing, and so

$$\begin{aligned} S(m) &\leq \sum_{t=0}^{\infty} \int_t^{t+1} \beta^t \min [s_t(m); \Delta_{\mathcal{X}}] ds \\ &\leq \frac{1}{\beta} \sum_{t=0}^{\infty} \int_t^{t+1} \beta^s \min [m(2K_F)^s; \Delta_{\mathcal{X}}] ds \\ &\leq \frac{1}{\beta} \int_0^{\infty} e^{-|\ln \beta|s} \min [me^{\ln(2K_F)s}; \Delta_{\mathcal{X}}] ds. \end{aligned} \quad (2.19)$$

Let  $t_\star$  be such that  $me^{\ln(2K_F)t_\star} = \Delta_{\mathcal{X}}$ , i.e.  $t_\star = \frac{\ln(\Delta_{\mathcal{X}}/m)}{\ln(2K_F)}$ . Then,

$$\begin{aligned} \int_0^{\infty} e^{-|\ln \beta|s} \min [me^{\ln(2K_F)s}; \Delta_{\mathcal{X}}] ds &\leq m \int_0^{t_\star} e^{\ln(2K_F\beta)s} ds + \Delta_{\mathcal{X}} \int_{t_\star}^{\infty} e^{\ln(\beta)s} ds \\ &\leq \frac{m}{\ln(2K_F\beta)} \left[ e^{\ln(2K_F\beta)t_\star} - 1 \right] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} e^{\ln(\beta)t_\star}. \end{aligned}$$

After substituting  $t_\star$  by its explicit value, we then obtain

$$\begin{aligned} \int_0^{\infty} e^{-|\ln \beta|s} \min [me^{\ln(2K_F)s}; \Delta_{\mathcal{X}}] ds &\leq \frac{m}{\ln(2K_F\beta)} \left[ \left( \frac{\Delta_{\mathcal{X}}}{m} \right)^{\frac{\ln(2K_F\beta)}{\ln(2K_F)}} - 1 \right] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} \left( \frac{\Delta_{\mathcal{X}}}{m} \right)^{\frac{\ln(\beta)}{\ln(2K_F)}} \\ &\leq \Delta_{\mathcal{X}} \left( \frac{1}{\ln(2K_F\beta)} - \frac{1}{\ln \beta} \right) \left( \frac{\Delta_{\mathcal{X}}}{m} \right)^{\frac{\ln(\beta)}{\ln(2K_F)}} - \frac{m}{\ln(2K_F\beta)} \\ &\leq \mathcal{O} \left( m^{\min \left[ 1, \frac{|\ln \beta|}{\ln(2K_F)} \right]} \right) = \mathcal{O} \left( M_N^{\min \left[ 1, \frac{|\ln \beta|}{\ln(2K_F)} \right]} \right). \end{aligned}$$

This, combined with (2), (2) and (2), concludes the proof.  $\square$

**Remark 2.4** If the Lipschitz constant in  $(\mathbf{HF}_{\text{lip}})$  satisfies  $\beta 2K_F < 1$ , then we can take  $\gamma = 1$  in the rate of convergence (2.6) of the optimal value.

In the particular case when  $F$  and  $f$  depend on the joint distribution  $\nu \in \mathcal{P}(\mathcal{X} \times A)$  only through its marginals on  $\mathcal{P}(\mathcal{X})$  and  $\mathcal{P}(A)$ , which is the usual framework considered in controlled mean-field dynamics, then a careful look in the above proof shows that the rate of convergence of the CMKV-MDP will be expressed in terms of

$$\tilde{M}_N := \max \left\{ \sup_{\mu \in \mathcal{P}(\mathcal{X})} \mathbb{E}[\mathcal{W}(\mu_N, \mu)], \sup_{\nu \in \mathcal{P}(A)} \mathbb{E}[\mathcal{W}_A(\nu_N, \nu)] \right\},$$

instead of  $M_N$  in (2), where here  $\mu_N$  (resp.  $\nu_N$ ) is the empirical measure associated to  $\mu$  (resp.  $\nu$ )  $\in \mathcal{P}(\mathcal{X})$  (resp.  $\mathcal{P}(A)$ ). From Lemma 2.1, the speed of convergence of  $\tilde{M}_N$  is faster than the one of  $M_N$ . For instance when  $\mathcal{X} \subset \mathbb{R}$ ,  $A \subset \mathbb{R}$ , then  $\tilde{M}_N = \mathcal{O}(N^{-1/2})$ , while  $M_N = \mathcal{O}(N^{-1/2} \log(1 + N))$ .

The propagation of chaos in mean-field control problems has been addressed in several papers, see [15], [5], [17], [7]. For instance, the recent paper [7] dealing with continuous time mean-field

diffusion processes, shows that the propagation of chaos holds when the optimization is performed over asymmetric open-loop controls where each individual control is allowed to depend upon all the past noises (including the noises from other individuals), in contrast to our model where the optimization is made over symmetric “individualized” open-loop controls (each control  $\alpha^i$  only depends upon the  $i$ -th individual’s noise and the common noise). However, rate of convergence is not provided in [7]. To the best of our knowledge, convergence rate has been obtained only in the continuous time case with feedback controls and finite horizon in Chapter 6, vol 2 of [5]. The originality of Theorem 2.1, besides providing a convergence rate in the individualized open-loop case, is that we are dealing with an infinite horizon. The above proof illustrates how two phenomena are involved in this result: for  $t \leq t_*$ , it is the regularity of  $F$  which ensures the closeness of the  $N$ -agent processes and the McKean-Vlasov processes, and thus the closeness of their rewards. As  $t$  becomes larger, because the horizon is infinite, the  $N$ -agent and McKean-Vlasov processes drift away from each other due to errors accumulation, but the compactness of  $\mathcal{X}$  prevents them from getting too far away from each other, and for  $t > t_*$ , the discounting factor  $\beta^t$ , by being smaller and smaller, makes the discounted rewards geometrically tend to 0 and be neglectable. What is remarkable is that these combined phenomena work together well enough to still ensure a fast convergence, even in infinite horizon, in  $\mathcal{O}(M_N^\gamma)$ , with an explicit order  $\gamma \leq 1$ .  $\square$

### 3 Lifted MDP on $\mathcal{P}(\mathcal{X})$

Theorem 2.1 justifies the CMKV-MDP as a macroscopic approximation of the  $N$ -agent MDP problem with mean-field interaction. Observe that the computation of the conditional gain, expected gain and optimal value of the CMKV-MDP in (2), only requires the variables associated to one agent, called representative agent. Therefore, we place ourselves in a reduced universe, dismissing other individuals variables, and renaming the representative agent’s initial information by  $\Gamma$ , initial state by  $\xi$ , idiosyncratic noise by  $(\varepsilon_t)_{t \in \mathbb{N}}$ . In the sequel, we shall denote by  $\mathcal{G} = \sigma(\Gamma)$  the  $\sigma$ -algebra generated by the random variable  $\Gamma$ , hence representing the initial information filtration, and by  $L^0(\mathcal{G}; \mathcal{X})$  the set of  $\mathcal{G}$ -measurable random variables valued in  $\mathcal{X}$ . We shall assume that the initial state  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ , which means that the policy has access to the agent’s initial state through the initial information filtration  $\mathcal{G}$ .

An open-loop control for the representative agent is a process  $\alpha$ , which is adapted to the filtration generated by  $(\Gamma, (\varepsilon_s)_{s \leq t}, (\varepsilon_s^0)_{s \leq t})_{t \in \mathbb{N}}$ , and associated to an open-loop policy by:  $\alpha_t = \alpha_t^\pi := \pi_t(\Gamma, (\varepsilon_s)_{s \leq t}, (\varepsilon_s^0)_{s \leq t})$  for some  $\pi \in \Pi_{OL}$ . We denote by  $\mathcal{A}$  the set of open-loop controls, and given  $\alpha \in \mathcal{A}$ ,  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ , the state process  $X = X^{\xi, \alpha}$  of the representative agent is governed by

$$X_{t+1} = F(X_t, \alpha_t, \mathbb{P}_{(X_t, \alpha_t)}^0, \varepsilon_{t+1}, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad X_0 = \xi. \quad (3.1)$$

For  $\alpha = \alpha^\pi$ ,  $\pi \in \Pi_{OL}$ , we write indifferently  $X^{\xi, \pi} = X^{\xi, \alpha}$ , and the expected gain  $V^\alpha = V^\pi$  equal to

$$V^\alpha(\xi) = \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \beta^t f(X_t, \alpha_t, \mathbb{P}_{(X_t, \alpha_t)}^0) \right],$$

where we stress the dependence upon the initial state  $\xi$ . The value function to the CMKV-MDP is then defined by

$$V(\xi) = \sup_{\alpha \in \mathcal{A}} V^\alpha(\xi), \quad \xi \in L^0(\mathcal{G}; \mathcal{X}).$$

Let us now show how one can lift the CMKV-MDP to a (classical) MDP on the space of probability measures  $\mathcal{P}(\mathcal{X})$ . We set  $\mathbb{F}^0$  as the filtration generated by the common noise  $\varepsilon^0$ . Given an open-loop control  $\alpha \in \mathcal{A}$ , and its state process  $X = X^{\xi, \alpha}$ , denote by  $\{\mu_t = \mathbb{P}_{X_t}^0, t \in \mathbb{N}\}$ , the random  $\mathcal{P}(\mathcal{X})$ -valued process, and notice from Proposition A.1 that  $(\mu_t)_t$  is  $\mathbb{F}^0$ -adapted. From (3), and recalling the pushforward measure notation, we have

$$\mu_{t+1} = F(\cdot, \cdot, \mathbb{P}_{(X_t, \alpha_t)}^0, \cdot, \varepsilon_{t+1}^0) \star (\mathbb{P}_{(X_t, \alpha_t)}^0 \otimes \lambda_\varepsilon), \quad a.s. \quad (3.2)$$

As the probability distribution  $\lambda_\varepsilon$  of the idiosyncratic noise is a fixed parameter, the above relation means that  $\mu_{t+1}$  only depends on  $\mathbb{P}_{(X_t, \alpha_t)}^0$  and  $\varepsilon_{t+1}^0$ . Moreover, by introducing the so-called *relaxed control* associated to the open-loop control  $\alpha$  as

$$\hat{\alpha}_t(x) = \mathcal{L}^0(\alpha_t | X_t = x), \quad t \in \mathbb{N},$$

which is valued in  $\hat{A}(\mathcal{X})$ , the set of probability kernels on  $\mathcal{X} \times A$  (see Lemma A.2), we see from Bayes formula that  $\mathbb{P}_{(X_t, \alpha_t)}^0 = \mu_t \cdot \hat{\alpha}_t$ . The dynamics relation (3) is then written as

$$\mu_{t+1} = \hat{F}(\mu_t, \hat{\alpha}_t, \varepsilon_{t+1}^0), \quad t \in \mathbb{N},$$

where the function  $\hat{F} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \times E^0 \rightarrow \mathcal{P}(\mathcal{X})$  is defined by

$$\hat{F}(\mu, \hat{\alpha}, e^0) = F(\cdot, \cdot, \mu \cdot \hat{\alpha}, \cdot, e^0) \star ((\mu \cdot \hat{\alpha}) \otimes \lambda_\varepsilon). \quad (3.3)$$

On the other hand, by the law of iterated conditional expectation, the expected gain can be written as

$$V^\alpha(\xi) = \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \beta^t \mathbb{E}^0 [f(X_t, \alpha_t, \mathbb{P}_{(X_t, \alpha_t)}^0)] \right],$$

with the conditional expectation term equal to

$$\mathbb{E}^0 [f(X_t, \alpha_t, \mathbb{P}_{(X_t, \alpha_t)}^0)] = \hat{f}(\mu_t, \hat{\alpha}_t),$$

where the function  $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \rightarrow \mathbb{R}$  is defined by

$$\hat{f}(\mu, \hat{\alpha}) = \int_{\mathcal{X} \times A} f(x, a, \mu \cdot \hat{\alpha})(\mu \cdot \hat{\alpha})(dx, da). \quad (3.4)$$

The above derivation suggests to consider a MDP with state space  $\mathcal{P}(\mathcal{X})$ , action space  $\hat{A}(\mathcal{X})$ , a state transition function  $\hat{F}$  as in (3), a discount factor  $\beta \in [0, 1)$ , and a reward function  $\hat{f}$  as in (3). A key point is to endow  $\hat{A}(\mathcal{X})$  with a suitable  $\sigma$ -algebra in order to have measurable functions  $\hat{F}$ ,  $\hat{f}$ , and  $\mathbb{F}^0$ -adapted process  $\hat{\alpha}$  valued in  $\hat{A}(\mathcal{X})$ , so that the MDP with characteristics  $(\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X}), \hat{F}, \hat{f}, \beta)$  is well-posed. This issue is investigated in the next sections, first in special cases, and then in general case by a suitable enlargement of the action space.

### 3.1 Case without common noise

When there is no common noise, the original state transition function  $F$  is defined from  $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E$  into  $\mathcal{X}$ , and the associated function  $\hat{F}$  is then defined from  $\mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$  into  $\mathcal{P}(\mathcal{X})$  by

$$\hat{F}(\mu, \hat{\alpha}) = F(\cdot, \cdot, \mu \cdot \hat{\alpha}, \cdot) \star ((\mu \cdot \hat{\alpha}) \otimes \lambda_\varepsilon).$$

In this case, we are simply reduced to a deterministic control problem on the state space  $\mathcal{P}(\mathcal{X})$  with dynamics

$$\mu_{t+1} = \hat{F}(\mu_t, \kappa_t), \quad t \in \mathbb{N}, \quad \mu_0 = \mu \in \mathcal{P}(\mathcal{X}),$$

controlled by  $\kappa = (\kappa_t)_{t \in \mathbb{N}} \in \hat{\mathcal{A}}$ , the set of deterministic sequences valued in  $\hat{A}(\mathcal{X})$ , and cumulated gain/value function:

$$\hat{V}^\kappa(\mu) = \sum_{t=0}^{\infty} \beta^t \hat{f}(\mu_t, \kappa_t), \quad \hat{V}(\mu) = \sup_{\kappa \in \hat{\mathcal{A}}} \hat{V}^\kappa(\mu), \quad \mu \in \mathcal{P}(\mathcal{X}),$$

where the bounded function  $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \rightarrow \mathbb{R}$  is defined as in (3). Notice that there are no measurability issues for  $\hat{F}$ ,  $\hat{f}$ , as the problem is deterministic and all the quantities defined above are well-defined.

We aim to prove the correspondence and equivalence between the MKV-MDP and the above deterministic control problem. From similar derivation as in (3)-(3) (by taking directly law under  $\mathbb{P}$  instead of  $\mathbb{P}^0$ ), we clearly see that for any  $\alpha \in \mathcal{A}$ ,  $V^\alpha(\xi) = \hat{V}^{\hat{\alpha}}(\mu)$ , with  $\mu = \mathcal{L}(\xi)$ , and  $\hat{\alpha} = \mathcal{R}_\xi(\alpha)$  where  $\mathcal{R}_\xi$  is the relaxed operator

$$\begin{aligned} \mathcal{R}_\xi : \mathcal{A} &\longrightarrow \hat{\mathcal{A}} \\ \alpha = (\alpha_t)_t &\longmapsto \hat{\alpha} = (\hat{\alpha}_t)_t : \hat{\alpha}_t(x) = \mathcal{L}(\alpha_t | X_t^{\xi, \alpha} = x), \quad t \in \mathbb{N}, \quad x \in \mathcal{X}. \end{aligned}$$

It follows that  $V(\xi) \leq \hat{V}(\mu)$ . In order to get the reverse inequality, we have to show that  $\mathcal{R}_\xi$  is surjective. Notice that this property is not always satisfied: for instance, when the  $\sigma$ -algebra generated by  $\xi$  is equal to  $\mathcal{G}$ , then for any  $\alpha \in \mathcal{A}$ ,  $\alpha_0$  is  $\sigma(\xi)$ -measurable at time  $t = 0$ , and thus  $\mathcal{L}(\alpha_0 | \xi)$  is a Dirac distribution, hence cannot be equal to an arbitrary probability kernel  $\kappa_0 = \hat{\alpha} \in \hat{A}(\mathcal{X})$ . We shall then make the following randomization hypothesis.

**Rand**( $\xi, \mathcal{G}$ ): There exists a uniform random variable  $U \sim \mathcal{U}([0, 1])$ , which is  $\mathcal{G}$ -measurable and independent of  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ .

**Remark 3.1** The randomization hypothesis **Rand**( $\xi, \mathcal{G}$ ) implies in particular that  $\Gamma$  is atomless, i.e.,  $\mathcal{G}$  is rich enough, and thus  $\mathcal{P}(\mathcal{X}) = \{\mathcal{L}(\zeta) : \zeta \in L^0(\mathcal{G}; \mathcal{X})\}$ . Furthermore, it means that there is extra randomness in  $\mathcal{G}$  besides  $\xi$ , so that one can freely randomize via the uniform random variable  $U$  the first action given  $\xi$  according to any probability kernel  $\hat{\alpha}$ . Moreover, one can extract from  $U$ , by standard separation of the decimals of  $U$  (see Lemma 2.21 in [14]), an i.i.d. sequence of uniform variables  $(U_t)_{t \in \mathbb{N}}$ , which are  $\mathcal{G}$ -measurable, independent of  $\xi$ , and can then be used to randomize the subsequent actions.  $\square$

**Theorem 3.1 (Correspondence in the no common noise case)**

Assume that **Rand**( $\xi, \mathcal{G}$ ) holds true. Then  $\mathcal{R}_\xi$  is surjective from  $\mathcal{A}$  into  $\hat{\mathcal{A}}$ , and we have  $V(\xi) = \hat{V}(\mu)$ , for  $\mu = \mathcal{L}(\xi)$ . Moreover, for  $\epsilon \geq 0$ , if  $\alpha^\epsilon \in \mathcal{A}$  is an  $\epsilon$ -optimal control for  $V(\xi)$ , then  $\mathcal{R}_\xi(\alpha^\epsilon) \in \hat{\mathcal{A}}$  is an  $\epsilon$ -optimal control for  $\hat{V}(\mu)$ , and conversely, if  $\hat{\alpha}^\epsilon \in \hat{\mathcal{A}}$  is an  $\epsilon$ -optimal control for  $\hat{V}(\mu)$ , then any  $\alpha^\epsilon \in \mathcal{R}_\xi^{-1}(\hat{\alpha}^\epsilon)$  is an  $\epsilon$ -optimal control for  $V(\xi)$ . Consequently, an optimal control for  $V(\xi)$  exists iff an optimal control for  $\hat{V}(\mu)$  exists.

**Proof.** In view of the above discussion, we only need to prove the surjectivity of  $\mathcal{R}_\xi$ . Fix a control  $\kappa \in \hat{\mathcal{A}}$  for the MDP on  $\mathcal{P}(\mathcal{X})$ . By Lemma 2.22 in [14], for all  $t \in \mathbb{N}$ , there exists a measurable function



$a_t : \mathcal{X} \times [0, 1] \rightarrow A$  such that  $\mathbb{P}_{a_t(x, U)} = \kappa_t(x)$ , for all  $x \in \mathcal{X}$ . It is then clear that the control  $\alpha$  defined recursively by  $\alpha_t := a_t(X_t^{\xi, \alpha}, U_t)$ , where  $(U_t)_t$  is an i.i.d. sequence of  $\mathcal{G}$ -measurable uniform variables independent of  $\xi$  under  $\mathbf{Rand}(\xi, \Gamma)$ , satisfies  $\mathcal{L}(\alpha_t | X_t^{\xi, \alpha} = x) = \kappa_t(x)$  (observing that  $U_t$  is independent of  $X_t^{\xi, \alpha}$ ), and thus  $\hat{\alpha} = \kappa$ , which proves the surjectivity of  $\mathcal{R}_\xi$ .  $\square$

**Remark 3.2** The above correspondence result shows in particular that the value function  $V$  of the MKV-MDP is law invariant, in the sense that it depends on its initial state  $\xi$  only via its probability law  $\mu = \mathcal{L}(\xi)$ , for  $\xi$  satisfying the randomization hypothesis.  $\square$

### 3.2 Case with finite state space $\mathcal{X}$ and with common noise

We consider the case with common noise but when the state space  $\mathcal{X}$  is finite, i.e., its cardinal  $\#\mathcal{X}$  is finite, equal to  $n$ .

In this case, one can identify  $\mathcal{P}(\mathcal{X})$  with the simplex  $\mathbb{S}^{n-1} = \{p = (p_i)_{i=1, \dots, n} \in [0, 1]^n : \sum_{i=1}^n p_i = 1\}$ , by associating any probability distribution  $\mu \in \mathcal{P}(\mathcal{X})$  to its weights  $(\mu(\{x\}))_{x \in \mathcal{X}} \in \mathbb{S}^{n-1}$ . We also identify the action space  $\hat{A}(\mathcal{X})$  with  $\mathcal{P}(A)^n$  by associating any probability kernel  $\hat{a} \in \hat{A}(\mathcal{X})$  to  $(\hat{a}(x))_{x \in \mathcal{X}} \in \mathcal{P}(A)^n$ , and thus  $\hat{A}(\mathcal{X})$  is naturally endowed with the product  $\sigma$ -algebra of the Wasserstein metric space  $\mathcal{P}(A)$ .

**Lemma 3.1** *Suppose that  $\#\mathcal{X} = n < \infty$ . Then,  $\hat{F}$  in (3) is a measurable function from  $\mathbb{S}^{n-1} \times \mathcal{P}(A)^n \times E^0$  into  $\mathbb{S}^{n-1}$ ,  $\hat{f}$  in (3) is a real-valued measurable function on  $\mathbb{S}^{n-1} \times \mathcal{P}(A)^n$ . Moreover, for any  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ , and  $\alpha \in \mathcal{A}$ , the  $\mathcal{P}(A)^n$ -valued process  $\hat{\alpha}$  defined by  $\hat{\alpha}_t(x) = \mathcal{L}^0(\alpha_t | X_t^{\xi, \alpha} = x)$ ,  $t \in \mathbb{N}$ ,  $x \in \mathcal{X}$ , is  $\mathbb{F}^0$ -adapted.*

**Proof.** By Lemma A.1, it is clear, by measurable composition, that we only need to prove that  $\Psi : (\mu, \hat{a}) \in (\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X})) \mapsto \mu \cdot \hat{a} \in \mathcal{P}(\mathcal{X} \times A)$  is measurable. However, in this finite state space case,  $\mu \cdot \hat{a}$  is here simply equal to  $\sum_{x \in \mathcal{X}} \mu(x) \hat{a}(x)$  and, thus  $\Psi$  is clearly measurable.  $\square$

In view of Lemma 3.1, the MDP with characteristics  $(\mathcal{P}(\mathcal{X}) \equiv \mathbb{S}^{n-1}, \hat{A}(\mathcal{X}) \equiv \mathcal{P}(A)^n, \hat{F}, \hat{f}, \beta)$  is well-posed. Let us then denote by  $\hat{\mathcal{A}}$  the set of  $\mathbb{F}^0$ -adapted processes valued in  $\mathcal{P}(A)^n$ , and given  $\kappa \in \hat{\mathcal{A}}$ , consider the controlled dynamics in  $\mathbb{S}^{n-1}$

$$\mu_{t+1} = \hat{F}(\mu_t, \kappa_t, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \mu_0 = \mu \in \mathbb{S}^{n-1}, \quad (3.5)$$

the associated expected gain and value function

$$\hat{V}^\kappa(\mu) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t \hat{f}(\mu_t, \kappa_t) \right], \quad \hat{V}(\mu) = \sup_{\kappa \in \hat{\mathcal{A}}} \hat{V}^\kappa(\mu). \quad (3.6)$$

We aim to prove the correspondence and equivalence between the CMKV-MDP and the MDP (3.2)-(3.2). From the derivation in (3)-(3) and by Lemma 3.1, we see that for any  $\alpha \in \mathcal{A}$ ,  $V^\alpha(\xi) = \hat{V}^{\hat{\alpha}}(\mu)$ , where  $\mu = \mathcal{L}(\xi)$ , and  $\hat{\alpha} = \mathcal{R}_\xi^0(\alpha)$  where  $\mathcal{R}_\xi^0$  is the relaxed operator

$$\begin{aligned} \mathcal{R}_\xi^0 : \mathcal{A} &\longrightarrow \hat{\mathcal{A}} \\ \alpha = (\alpha_t)_t &\longmapsto \hat{\alpha} = (\hat{\alpha}_t)_t : \hat{\alpha}_t(x) = \mathcal{L}^0(\alpha_t | X_t^{\xi, \alpha} = x), \quad t \in \mathbb{N}, x \in \mathcal{X}. \end{aligned} \quad (3.7)$$

It follows that  $V(\xi) \leq \hat{V}(\mu)$ . In order to get the reverse inequality from the surjectivity of  $\mathcal{R}_\xi^0$ , we need again as in the no common noise case to make some randomization hypothesis. It turns out that when  $\mathcal{X}$  is finite, this randomization hypothesis is simply reduced to the atomless property of  $\Gamma$ .

**Lemma 3.2** *Assume that  $\Gamma$  is atomless, i.e.,  $\mathcal{G}$  is rich enough. Then, any  $\xi \in L^0(\mathcal{G}; \mathcal{X})$  taking a countable number of values, satisfies  $\mathbf{Rand}(\xi, \Gamma)$ .*

**Proof.** Let  $S$  be a countable set s.t.  $\xi \in S$  a.s., and  $\mathbb{P}[\xi = x] > 0$  for all  $x \in S$ . Fix  $x \in S$  and denote by  $\mathbb{P}_x$  the probability “knowing  $\xi = x$ ”, i.e.,  $\mathbb{P}_x[B] := \frac{\mathbb{P}[B, \xi=x]}{\mathbb{P}[\xi=x]}$ , for all  $B \in \mathcal{F}$ . It is clear that, endowing  $\Omega$  with this probability,  $\Gamma$  is still atomless, and so there exists a  $\mathcal{G}$ -measurable random variable  $U_x$  that is uniform under  $\mathbb{P}_x$ . Then, the random variable  $U := \sum_{x \in S} U_x \mathbf{1}_{\xi=x}$  is a  $\mathcal{G}$ -measurable uniform random variable under  $\mathbb{P}_x$  for all  $x \in S$ , which implies that it is a uniform variable under  $\mathbb{P}$ , independent of  $\xi$ .  $\square$

**Theorem 3.2 (Correspondance with the MDP on  $\mathcal{P}(\mathcal{X})$  in the  $\mathcal{X}$  finite case)**

*Assume that  $\mathcal{G}$  is rich enough. Then  $\mathcal{R}_\xi^0$  is surjective from  $\mathcal{A}$  into  $\hat{\mathcal{A}}$ , and  $V(\xi) = \hat{V}(\mu)$ , for any  $\mu \in \mathcal{P}(\mathcal{X})$ ,  $\xi \in L^0(\mathcal{G}; \mathcal{X})$  s.t.  $\mu = \mathcal{L}(\xi)$ . Moreover, for  $\epsilon \geq 0$ , if  $\alpha^\epsilon \in \mathcal{A}$  is an  $\epsilon$ -optimal control for  $V(\xi)$ , then  $\mathcal{R}_\xi^0(\alpha^\epsilon) \in \hat{\mathcal{A}}$  is an  $\epsilon$ -optimal control for  $\hat{V}(\mu)$ . Conversely, if  $\hat{\alpha}^\epsilon \in \hat{\mathcal{A}}$  is an  $\epsilon$ -optimal control for  $\hat{V}(\mu)$ , then any  $\alpha^\epsilon \in (\mathcal{R}_\xi^0)^{-1}(\hat{\alpha}^\epsilon)$  is an  $\epsilon$ -optimal control for  $V(\xi)$ . Consequently, an optimal control for  $V(\xi)$  exists iff an optimal control for  $\hat{V}(\mu)$  exists.*

**Proof.** From the derivation in (3.2)-(3.2), we only need to prove the surjectivity of  $\mathcal{R}_\xi^0$ . Fix  $\kappa \in \hat{\mathcal{A}}$  and let  $\pi_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$  be such that  $\kappa_t = \pi_t((\varepsilon_s^0)_{s \leq t})$ . As  $\mathcal{X}$  is finite, by definition of the  $\sigma$ -algebra on  $\hat{A}(\mathcal{X})$ ,  $\pi_t$  can be seen as a measurable function in  $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$ . Let  $\phi \in L^0(A, \mathbb{R})$  be an embedding as in Lemma C.2. By Lemma A.1, we know that  $\phi \star \pi_t$  is in  $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(\mathbb{R}))$ . Given  $m \in \mathcal{P}(\mathbb{R})$  we denote by  $F_m^{-1}$  the generalized inverse of its distribution function, and it is known that the mapping  $m \in (\mathcal{P}(\mathbb{R}), \mathcal{W}) \mapsto F_m^{-1} \in (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1)$  is an isometry and is thus measurable. Therefore,  $F_{\phi \star \pi_t}^{-1}$  is in  $L^0((E^0)^t \times \mathcal{X}; (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1))$ . Finally, the mapping  $(f, u) \in (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1) \times ([0, 1], \mathcal{B}([0, 1])) \mapsto f(u) \in (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  is measurable, since it is the limit of the sequence  $n \sum_{i \in \mathbb{Z}} \mathbf{1}_{[\frac{i+1}{n}, \frac{i+2}{n})}(u) \int_{\frac{i}{n}}^{\frac{i+1}{n}} f(y) dy$  when  $n \rightarrow \infty$ . Therefore, the mapping

$$\begin{aligned} \alpha_t : (E^0)^t \times \mathcal{X} \times [0, 1] &\longrightarrow A \\ ((\varepsilon_s^0)_{s \leq t}, x, u) &\longmapsto \phi^{-1} \circ F_{\phi \star \pi_t((\varepsilon_s^0)_{s \leq t}, x)}^{-1}(u) \end{aligned}$$

is measurable. We thus define, by induction,  $\alpha_t := \alpha_t((\varepsilon_s^0)_{s \leq t}, X_t^{\xi, \alpha}, U_t)$ . By construction and by the generalized inverse simulation method, it is clear that  $\hat{\alpha}_t = \kappa_t$ .  $\square$

**Remark 3.3** We point out that when both state space  $\mathcal{X}$  and action space  $A$  are finite, equipped with the metrics  $d(x, x') := \mathbf{1}_{x \neq x'}$ ,  $x, x' \in \mathcal{X}$  and  $d_A(a, a') := \mathbf{1}_{a \neq a'}$ ,  $a, a' \in A$ , the transition function  $\hat{F}$  and reward function  $\hat{f}$  of the lifted MDP on  $\mathcal{P}(\mathcal{X})$  inherits the Lipschitz condition  $(\mathbf{HF}_{\text{lip}})$  and  $(\mathbf{Hf}_{\text{lip}})$  used for the propagation of chaos. Indeed, it is known that the Wasserstein distance obtained from  $d$  (resp.  $d_A$ ) coincides with twice the total variation distance, and thus to the  $\mathbb{L}^1$  distance when naturally embedding  $\mathcal{P}(\mathcal{X})$  (resp.  $\mathcal{P}(A)$ ) in  $[0, 1]^{\#\mathcal{X}}$  (resp.  $[0, 1]^{\#A}$ ). Thus, embedding  $\hat{A}(\mathcal{X})$  in  $\mathcal{M}_{\#\mathcal{X}, \#A}([0, 1])$ , the set of  $\#\mathcal{X} \times \#A$  matrices with coefficients valued in  $[0, 1]$ , we have

$$\|\hat{F}(\mu, \hat{a}, e^0), \hat{F}(\nu, \hat{a}', e^0)\|_1 \leq (1 + K_F)(2\|\mu - \nu\|_1 + \sup_{x \in \mathcal{X}} \|\hat{a}_{x, \cdot} - \hat{a}'_{x, \cdot}\|_1).$$

We obtain a similar property for  $f$ . In other words, lifting the CMKV-MDP not only turns it into an MDP, but also its state and action spaces  $[0, 1]^{\#\mathcal{X}}$  and  $[0, 1]^{\#\mathcal{X} \times \#A}$  are very standard, and its dynamic

and reward are Lipschitz functions with factors of the order of  $K_F$  and  $K_f$  according to the norm  $\|\cdot\|_1$ . Thus, due to the standard nature of this MDP, most MDP algorithms can be applied and their speed will be simply expressed in terms of the original parameters of the CMKV-MDP,  $K_F$  and  $K_f$ .  $\square$

**Remark 3.4** As in the no common noise case, the correspondence result in the finite state space case for  $\mathcal{X}$  shows notably that the value function of the CMKV-MDP is law-invariant.

The general case (common noise and continuous state space  $\mathcal{X}$ ) raises multiple issues for establishing the equivalence between CMKV-MDP and the lifted MDP on  $\mathcal{P}(\mathcal{X})$ . First, we have to endow the action space  $\hat{A}(\mathcal{X})$  with a suitable  $\sigma$ -algebra for the lifted MDP to be well-posed: on the one hand, this  $\sigma$ -algebra has to be large enough to make the functions  $\hat{F} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \times E^0 \rightarrow \mathcal{P}(\mathcal{X})$  and  $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \rightarrow \mathbb{R}$  measurable, and on the other hand, it should be small enough to make the process  $\hat{\alpha} = \mathcal{R}_\xi^0(\alpha)$   $\mathbb{F}^0$ -adapted for any control  $\alpha \in \mathcal{A}$  in the CMKV-MDP. Beyond the well-posedness issue of the lifted MDP, the second important concern is the surjectivity of the relaxed operator  $\mathcal{R}_\xi^0$  from  $\mathcal{A}$  into  $\hat{\mathcal{A}}$ . Indeed, if we try to adapt the proof of Theorem 3.2 to the case of a continuous state space  $\mathcal{X}$ , the issue is that we cannot in general equip  $\hat{A}(\mathcal{X})$  with a  $\sigma$ -algebra such that  $L^0((E^0)^t; \hat{A}(\mathcal{X})) = L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$ , and thus we cannot see  $\pi_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$  as an element of  $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$ , which is crucial because the control  $\alpha$  (such that  $\hat{\alpha} = \kappa$ ) is defined with  $\alpha_t$  explicitly depending upon  $\pi_t((\varepsilon_s^0)_{s \leq t}, X_t)$ .

In the next section, we shall fix these measurability issues in the general case, and prove the correspondence between the CMKV-MDP and a general lifted MDP on  $\mathcal{P}(\mathcal{X})$ .  $\square$

## 4 General case and Bellman fixed point equation in $\mathcal{P}(\mathcal{X})$

We address the general case with common noise and possibly continuous state space  $\mathcal{X}$ , and our aim is to state the correspondence of the CMKV-MDP with a suitable lifted MDP on  $\mathcal{P}(\mathcal{X})$  associated to a Bellman fixed point equation, characterizing the value function, and obtain as a by-product an  $\epsilon$ -optimal control. We proceed as follows:

- (i) We first introduce a well-posed lifted MDP on  $\mathcal{P}(\mathcal{X})$  by enlarging the action space to  $\mathcal{P}(\mathcal{X} \times A)$ , and call  $\tilde{V}$  the corresponding value function, which satisfies:  $V(\xi) \leq \tilde{V}(\mu)$ , for  $\mu = \mathcal{L}(\xi)$ .
- (ii) We then consider the Bellman equation associated to this well-posed lifted MDP on  $\mathcal{P}(\mathcal{X})$ , which admits a unique fixed point, called  $V^*$ .
- (iii) Under the randomization hypothesis for  $\xi$ , we show the existence of an  $\epsilon$ -randomized feedback policy, which yields both an  $\epsilon$ -randomized feedback control for the CMKV-MDP and an  $\epsilon$ -optimal feedback control for  $\tilde{V}$ . This proves that  $V(\xi) = \tilde{V}(\mu) = V^*(\mu)$ , for  $\mu = \mathcal{L}(\xi)$ .
- (iv) Under the condition that  $\mathcal{G}$  is rich enough, we conclude that  $V$  is law-invariant and is equal to  $\tilde{V} = V^*$ , hence satisfies the Bellman equation.

Finally, we show how to compute from the Bellman equation by value or policy iteration approximate optimal strategy and value function.

### 4.1 A general lifted MDP on $\mathcal{P}(\mathcal{X})$

We start again from the relation (3) describing the evolution of  $\mu_t = \mathbb{P}_{X_t}^0$ ,  $t \in \mathbb{N}$ , for a state process  $X_t = X_t^{\xi, \alpha}$  controlled by  $\alpha \in \mathcal{A}$ :

$$\mu_{t+1} = F(\cdot, \cdot, \mathbb{P}_{(X_t, \alpha_t)}^0, \cdot, \varepsilon_{t+1}^0) \star (\mathbb{P}_{(X_t, \alpha_t)}^0 \otimes \lambda_\varepsilon), \quad a.s. \quad (4.1)$$

Now, instead of disintegrating as in Section 3, the conditional law of the pair  $(X_t, \alpha_t)$ , as  $\mathbb{P}_{(X_t, \alpha_t)}^0 = \mu_t \cdot \hat{\alpha}_t$  where  $\hat{\alpha} = \mathcal{R}_\xi^0(\alpha)$  is the relaxed control in (3.2), we directly consider the control process  $\alpha_t = \mathbb{P}_{(X_t, \alpha_t)}^0$ ,  $t \in \mathbb{N}$ , which is  $\mathbb{F}^0$ -adapted (see Proposition A.1), and valued in the space of probability measures  $\mathbf{A} := \mathcal{P}(\mathcal{X} \times A)$ , naturally endowed with the  $\sigma$ -algebra of its Wasserstein metric. Notice that this  $\mathbf{A}$ -valued control  $\alpha$  obtained from the CMKV-MDP has to satisfy by definition the marginal constraint  $\text{pr}_1 \star \alpha_t = \mu_t$  at any time  $t$ . In order to tackle this marginal constraint, we shall rely on the following coupling results.

**Lemma 4.1 (Measurable coupling)**

There exists a measurable function  $\zeta \in L^0(\mathcal{P}(\mathcal{X})^2 \times \mathcal{X} \times [0, 1]; \mathcal{X})$  s.t. for any  $(\mu, \mu') \in \mathcal{P}(\mathcal{X})$ , and if  $\xi \sim \mu$ , then

- $\zeta(\mu, \mu', \xi, U) \sim \mu'$ , where  $U$  is an uniform random variable independent of  $\xi$ .
- (i) When  $\mathcal{X} \subset \mathbb{R}$ :

$$\mathbb{E}[d(\xi, \zeta(\mu, \mu', \xi, U))] = \mathcal{W}(\mu, \mu').$$

(ii) In general when  $\mathcal{X}$  Polish:  $\forall \varepsilon > 0, \exists \eta > 0$  s.t.

$$\mathcal{W}(\mu, \mu') < \eta \Rightarrow \mathbb{E}[d(\xi, \zeta(\mu, \mu', \xi, U))] < \varepsilon.$$

**Proof.** See Appendix C. □

**Remark 4.1** Lemma 4.1 can be seen as a measurable version of the well-known coupling result in optimal transport, which states that given  $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ , there exists  $\xi$  and  $\xi'$  random variables with  $\mathcal{L}(\xi) = \mu, \mathcal{L}(\xi') = \mu'$  such that  $\mathcal{W}(\mu, \mu') = \mathbb{E}[d(\xi, \xi')]$ . A similar measurable optimal coupling is proved in [9] under the assumption that there exists a transfer function realizing an optimal coupling between  $\mu$  and  $\mu'$ . However, such transfer function does not always exist, for instance when  $\mu$  has atoms but not  $\mu'$ . Lemma 4.1 builds a measurable coupling without making such assumption (essentially using the uniform variable  $U$  to randomize when  $\mu$  has atoms). □

From the measurable coupling function  $\zeta$  as in Lemma 4.1, we define the coupling projection  $\mathbf{p} : \mathcal{P}(\mathcal{X}) \times \mathbf{A} \rightarrow \mathbf{A}$  by

$$\mathbf{p}(\mu, \mathbf{a}) = \mathcal{L}(\zeta(\text{pr}_1 \star \mathbf{a}, \mu, \xi', U), \alpha_0), \quad \mu \in \mathcal{P}(\mathcal{X}), \mathbf{a} \in \mathbf{A},$$

where  $(\xi', \alpha_0) \sim \mathbf{a}$ , and  $U$  is a uniform random variable independent of  $\xi'$ .

**Lemma 4.2 (Measurable coupling projection)**

The coupling projection  $\mathbf{p}$  is a measurable function from  $\mathcal{P}(\mathcal{X}) \times \mathbf{A}$  into  $\mathbf{A}$ , and for all  $(\mu, \mathbf{a}) \in \mathcal{P}(\mathcal{X}) \times \mathbf{A}$ :

$$\text{pr}_1 \star \mathbf{p}(\mu, \mathbf{a}) = \mu, \quad \text{and if } \text{pr}_1 \star \mathbf{a} = \mu, \text{ then } \mathbf{p}(\mu, \mathbf{a}) = \mathbf{a}. \quad (4.2)$$

**Proof.** By construction, it is clear that  $\zeta(\mu, \mu, \xi, U) = \xi$ , and so relation (4.2) is obvious. The only result that is not trivial is the measurability of  $\mathbf{p}$ . Observe that  $\mathbf{p}(\mu, \mathbf{a}) = g(\mu, \mathbf{a}, \cdot, \cdot, \cdot) \star (\mathbf{a} \otimes \mathcal{U}([0, 1]))$  where  $g$  is the measurable function

$$\begin{aligned} g : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times A) \times \mathcal{X} \times A \times [0, 1] &\longrightarrow \mathcal{X} \times A \\ (\mu, \mathbf{a}, x, a, u) &\longmapsto (\zeta(\text{pr}_1 \star \mathbf{a}, \mu, x, u), a) \end{aligned}$$

We thus conclude by Lemma A.1.  $\square$

By using this coupling projection  $\mathbf{p}$ , we see that the dynamics (4.1) can be written as

$$\mu_{t+1} = \tilde{F}(\mu_t, \boldsymbol{\alpha}_t, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad (4.3)$$

where the function  $\tilde{F} : \mathcal{P}(\mathcal{X}) \times \mathbf{A} \times E^0 \rightarrow \mathcal{P}(\mathcal{X})$  defined by

$$\tilde{F}(\mu, \mathbf{a}, e^0) = F(\cdot, \cdot, \mathbf{p}(\mu, \mathbf{a}), \cdot, e^0) \star (\mathbf{p}(\mu, \mathbf{a}) \otimes \lambda_\varepsilon),$$

is clearly measurable. Let us also define the measurable function  $\tilde{f} : \mathcal{P}(\mathcal{X}) \times \mathbf{A} \rightarrow \mathbb{R}$  by

$$\tilde{f}(\mu, \mathbf{a}) = \int_{\mathcal{X} \times A} f(x, a, \mathbf{p}(\mu, \mathbf{a})) \mathbf{p}(\mu, \mathbf{a})(dx, da).$$

The MDP with characteristics  $(\mathcal{P}(\mathcal{X}), \mathbf{A} = \mathcal{P}(\mathcal{X} \times A), \tilde{F}, \tilde{f}, \beta)$  is then well-posed. Let us then denote by  $\mathcal{A}$  the set of  $\mathbb{F}^0$ -adapted processes valued in  $\mathbf{A}$ , and given an open-loop control  $\boldsymbol{\nu} \in \mathcal{A}$ , consider the controlled dynamics

$$\mu_{t+1} = \tilde{F}(\mu_t, \boldsymbol{\nu}_t, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad \mu_0 = \mu \in \mathcal{P}(\mathcal{X}), \quad (4.4)$$

with associated expected gain/value function

$$\tilde{V}^\nu(\mu) = \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \beta^t \tilde{f}(\mu_t, \boldsymbol{\nu}_t) \right], \quad \tilde{V}(\mu) = \sup_{\boldsymbol{\nu} \in \mathcal{A}} \tilde{V}^\nu(\mu). \quad (4.5)$$

Given  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ , and  $\alpha \in \mathcal{A}$ , we set  $\boldsymbol{\alpha} = \mathcal{L}_\xi^0(\alpha)$ , where  $\mathcal{L}_\xi^0$  is the lifted operator

$$\begin{aligned} \mathcal{L}_\xi^0 : \mathcal{A} &\longrightarrow \mathcal{A} \\ \alpha = (\alpha_t)_t &\longmapsto \boldsymbol{\alpha} = (\boldsymbol{\alpha}_t)_t : \boldsymbol{\alpha}_t = \mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \quad t \in \mathbb{N}. \end{aligned}$$

By construction from (4.1), we see that  $\mu_t = \mathbb{P}_{X_t^{\xi, \alpha}}^0$ ,  $t \in \mathbb{N}$ , follows the dynamics (4.1) with the control  $\boldsymbol{\nu} = \mathcal{L}_\xi^0(\alpha) \in \mathcal{A}$ . Moreover, by the law of iterated conditional expectation, and the definition of  $\tilde{f}$ , the expected gain of the CMKV-MDP can be written as

$$\begin{aligned} V^\alpha(\xi) &= \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \beta^t \mathbb{E}^0 [f(X_t^{\xi, \alpha}, \alpha_t, \mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0)] \right] \\ &= \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \beta^t \tilde{f}(\mathbb{P}_{X_t^{\xi, \alpha}}^0, \boldsymbol{\alpha}_t) \right] = \tilde{V}^\alpha(\mu), \quad \text{with } \mu = \mathcal{L}(\xi). \end{aligned} \quad (4.6)$$

It follows that  $V(\xi) \leq \tilde{V}(\mu)$ , for  $\mu = \mathcal{L}(\xi)$ . Our goal is to prove the equality, which implies in particular that  $V$  is law-invariant, and to obtain as a by-product the corresponding Bellman fixed point equation that characterizes analytically the solution to the CMKV-MDP.

## 4.2 Bellman fixed point on $\mathcal{P}(\mathcal{X})$

We derive and study the Bellman equation corresponding to the general lifted MDP (4.1)-(4.1) on  $\mathcal{P}(\mathcal{X})$ .

By defining this MDP on the canonical space  $(E^0)^\mathbb{N}$ , we identify  $\varepsilon^0$  with the canonical identity function in  $(E^0)^\mathbb{N}$ , and  $\varepsilon_t^0$  with the  $t$ -th projection in  $(E^0)^\mathbb{N}$ . We also denote by  $\theta : (E^0)^\mathbb{N} \rightarrow (E^0)^\mathbb{N}$  the

shifting operator, defined by  $\theta((e_t^0)_{t \in \mathbb{N}}) = (e_{t+1}^0)_{t \in \mathbb{N}}$ . Via this identification, an open-loop control  $\boldsymbol{\nu} \in \mathcal{A}$  is a sequence  $(\boldsymbol{\nu}_t)_t$  where  $\boldsymbol{\nu}_t$  is a measurable function from  $(E^0)^t$  into  $\mathbf{A}$ , with the convention that  $\boldsymbol{\nu}_0$  is simply a constant in  $\mathbf{A}$ . Given  $\boldsymbol{\nu} \in \mathcal{A}$ , and  $e^0 \in E^0$ , we define  $\tilde{\boldsymbol{\nu}}^{e^0} := (\tilde{\boldsymbol{\nu}}_t^{e^0})_t \in \mathcal{A}$ , where  $\tilde{\boldsymbol{\nu}}_t^{e^0}(\cdot) := \boldsymbol{\nu}_{t+1}(e^0, \cdot)$ ,  $t \in \mathbb{N}$ . Given  $\mu \in \mathcal{P}(\mathcal{X})$ , and  $\boldsymbol{\nu} \in \mathcal{A}$ , we denote by  $(\mu_t^{\mu, \boldsymbol{\nu}})_t$  the solution to (4.1) on the canonical space, which satisfies the flow property

$$(\mu_{t+1}^{\mu, \boldsymbol{\nu}}, \boldsymbol{\nu}_{t+1}) \equiv (\mu_t^{\mu_1^{\mu, \boldsymbol{\nu}}, \tilde{\boldsymbol{\nu}}^{\varepsilon_1^0}(\theta(\varepsilon^0))}, \tilde{\boldsymbol{\nu}}_t^{\varepsilon_1^0}(\theta(\varepsilon^0))), \quad t \in \mathbb{N}.$$

where  $\equiv$  denotes the equality between functions on the canonical space. Given that  $\varepsilon_1^0 \perp \theta(\varepsilon^0) \stackrel{d}{=} \varepsilon^0$ , we obtain that the expected gain of this MDP in (4.1) satisfies the relation

$$\tilde{V}^{\boldsymbol{\nu}}(\mu) = \tilde{f}(\mu, \boldsymbol{\nu}_0) + \beta \mathbb{E} \left[ \tilde{V}^{\tilde{\boldsymbol{\nu}}^{\varepsilon_1^0}}(\mu_1^{\mu, \boldsymbol{\nu}}) \right]. \quad (4.7)$$

Let us denote by  $L^\infty(\mathcal{P}(\mathcal{X}))$  the set of bounded real-valued functions on  $\mathcal{P}(\mathcal{X})$ , and by  $L_m^\infty(\mathcal{P}(\mathcal{X}))$  the subset of measurable functions in  $L^\infty(\mathcal{P}(\mathcal{X}))$ . We then introduce the Bellman ‘‘operator’’  $\mathcal{T} : L_m^\infty(\mathcal{P}(\mathcal{X})) \rightarrow L^\infty(\mathcal{P}(\mathcal{X}))$  defined for any  $W \in L_m^\infty(\mathcal{P}(\mathcal{X}))$  by:

$$[\mathcal{T}W](\mu) := \sup_{\mathbf{a} \in \mathbf{A}} \left\{ \tilde{f}(\mu, \mathbf{a}) + \beta \mathbb{E} [W(\tilde{F}(\mu, \mathbf{a}, \varepsilon_1^0))] \right\}, \quad \mu \in \mathcal{P}(\mathcal{X}). \quad (4.8)$$

Notice that the sup can a priori lead to a non measurable function  $\mathcal{T}W$ . This Bellman operator is consistent with the lifted MDP derived in Section 3, with characteristics  $(\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X}), \hat{F}, \hat{f}, \beta)$ , although this MDP is not always well-posed. Indeed, its corresponding Bellman operator is well-defined as it only involves the random variable  $\varepsilon_1^0$  at time 1, hence only requires the measurability of  $e^0 \mapsto \hat{F}(\mu, \hat{a}, e^0)$ , for any  $(\mu, \hat{a}) \in \mathbb{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$  (which holds true), and it turns out that it coincides with  $\mathcal{T}$ .

**Proposition 4.1** *For any  $W \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ , and  $\mu \in \mathcal{P}(\mathcal{X})$ , we have*

$$[\mathcal{T}W](\mu) = \sup_{\hat{a} \in \hat{A}(\mathcal{X})} [\hat{\mathcal{T}}^{\hat{a}}W](\mu) = \sup_{\mathbf{a} \in L^0(\mathcal{X} \times [0, 1]; \mathbf{A})} [\mathbb{T}^{\mathbf{a}}W](\mu), \quad (4.9)$$

where  $\hat{\mathcal{T}}^{\hat{a}}$  and  $\mathbb{T}^{\mathbf{a}}$  are the operators defined on  $L^\infty(\mathcal{P}(\mathcal{X}))$  by

$$\begin{aligned} [\hat{\mathcal{T}}^{\hat{a}}W](\mu) &= \hat{f}(\mu, \hat{a}) + \beta \mathbb{E} [W(\hat{F}(\mu, \hat{a}, \varepsilon_1^0))], \\ [\mathbb{T}^{\mathbf{a}}W](\mu) &= \mathbb{E} \left[ f(\xi, \mathbf{a}(\xi, U), \mathcal{L}(\xi, \mathbf{a}(\xi, U))) + \beta W(\mathbb{P}_{F(\xi, \mathbf{a}(\xi, U), \mathcal{L}(\xi, \mathbf{a}(\xi, U)), \varepsilon_1, \varepsilon_1^0)}^0) \right], \end{aligned} \quad (4.10)$$

for any  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$  (it is clear that the right-hand side in (4.1) does not depend on the choice of such  $(\xi, U)$ ). Moreover, we have

$$[\mathcal{T}W](\mu) = \sup_{\alpha_0 \in L^0(\Omega; \mathbf{A})} \mathbb{E} \left[ f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)) + \beta W(\mathbb{P}_{F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)), \varepsilon_1, \varepsilon_1^0}^0) \right]. \quad (4.11)$$

**Proof.** Fix  $W \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ , and  $\mu \in \mathcal{P}(\mathcal{X})$ . Let  $\mathbf{a}$  be arbitrary in  $\mathbf{A}$ . Since  $\mathbf{p}(\mu, \mathbf{a})$  has first marginal equal to  $\mu$ , there exists by assertion 3 in Lemma A.2 a probability kernel  $\hat{a} \in \hat{A}(\mathcal{X})$  such that  $\mathbf{p}(\mu, \mathbf{a}) = \mu \cdot \hat{a}$ . Therefore,  $\tilde{F}(\mu, \mathbf{a}, e^0) = \hat{F}(\mu, \hat{a}, e^0)$ ,  $\tilde{f}(\mu, \mathbf{a}) = \hat{f}(\mu, \hat{a})$ , which implies that  $[\mathcal{T}W](\mu) \leq \sup_{\hat{a} \in \hat{A}(\mathcal{X})} [\hat{\mathcal{T}}^{\hat{a}}W](\mu) =: \mathbb{T}^1$ .

Let us consider the operator  $\mathcal{R}$  defined by

$$\begin{aligned} \mathcal{R} : L^0(\mathcal{X} \times [0, 1]; \mathbf{A}) &\longrightarrow \hat{A}(\mathcal{X}) \\ \mathbf{a} &\longmapsto \hat{a} : \hat{a}(x) = \mathcal{L}(\mathbf{a}(x, U)), \quad x \in \mathcal{X}, U \sim \mathcal{U}([0, 1]), \end{aligned}$$

and notice that it is surjective from  $L^0(\mathcal{X} \times [0, 1]; A)$  into  $\hat{A}(\mathcal{X})$ , by Lemma 2.22 in [14]. By noting that for any  $\mathbf{a} \in L^0(\mathcal{X} \times [0, 1]; A)$ , and  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ , we have  $\mathcal{L}(\xi, \mathbf{a}(\xi, U)) = \mu \cdot \mathcal{R}(\mathbf{a})$ , it follows that  $[\mathbb{T}^{\mathbf{a}}W](\mu) = [\hat{\mathcal{T}}^{\mathcal{R}(\mathbf{a})}W](\mu)$ . Since  $\mathcal{R}$  is surjective, this yields  $\mathbb{T}^1 = \sup_{\mathbf{a} \in L^0(\mathcal{X} \times [0, 1]; A)} [\mathbb{T}^{\mathbf{a}}W](\mu) =: \mathbb{T}^2$ .

Denote by  $\mathbb{T}^3$  the right-hand-side in (4.1). It is clear that  $\mathbb{T}^2 \leq \mathbb{T}^3$ . Conversely, let  $\alpha_0 \in L^0(\Omega; A)$ . We then set  $\mathbf{a} = \mathcal{L}(\xi, \alpha_0) \in \mathcal{P}(\mathcal{X} \times A)$ , and notice that the first marginal of  $\mathbf{a}$  is  $\mu$ . Thus,  $\mathbf{p}(\mu, \mathbf{a}) = \mathcal{L}(\xi, \alpha_0)$ , and so

$$\begin{aligned}\tilde{f}(\mu, \mathbf{a}) &= \int_{\mathcal{X} \times A} f(x, a, \mathbf{p}(\mu, \mathbf{a})) \mathbf{p}(\mu, \mathbf{a})(dx, da) = \mathbb{E}[f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0))] \\ \tilde{F}(\mu, \mathbf{a}, \varepsilon_1^0) &= F(\cdot, \cdot, \mathbf{p}(\mu, \mathbf{a}), \cdot, \varepsilon_1^0) \star (\mathbf{p}(\mu, \mathbf{a}) \otimes \lambda_\varepsilon) = \mathbb{P}_{F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1, \varepsilon_1^0)}^0.\end{aligned}$$

We deduce that  $\mathbb{T}^3 \leq [\mathcal{T}W](\mu)$ , which gives finally the equalities (4.1) and (4.1).  $\square$

We state the basic properties of the Bellman operator  $\mathcal{T}$ .

**Proposition 4.2** *Assume that  $(\mathbf{H}_{\text{ip}})$  holds true. (i) The operator  $\mathcal{T}$  is monotone increasing: for  $W_1, W_2 \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ , if  $W_1 \leq W_2$ , then  $\mathcal{T}W_1 \leq \mathcal{T}W_2$ . (ii) Furthermore, it is contracting on  $L_m^\infty(\mathcal{P}(\mathcal{X}))$  with Lipschitz factor  $\beta$ , and admits a unique fixed point in  $L_m^\infty(\mathcal{P}(\mathcal{X}))$ , denoted by  $V^*$ , hence solution to:*

$$V^* = \mathcal{T}V^*.$$

(iii)  $V^*$  is  $\gamma$ -Hölder, with  $\gamma = \min\left(1, \frac{|\ln \beta|}{\ln(2K_F)}\right)$ , i.e. there exists some positive constant  $K_\star$  (depending only on  $K_F, K_f, \beta$ , and explicit in the proof), such that

$$|V^*(\mu) - V^*(\mu')| \leq K_\star \mathcal{W}(\mu, \mu')^\gamma, \quad \forall \mu, \mu' \in \mathcal{P}(\mathcal{X}).$$

**Proof.** (i) The monotonicity of  $\mathcal{T}$  is shown by standard arguments.

(ii) The  $\beta$ -contraction property of  $\mathcal{T}$  is also obtained by standard arguments. Let us now prove by induction that the iterative sequence  $V_{n+1} = \mathcal{T}V_n$ , with  $V_0 \equiv 0$  is well defined and such that

$$|V_n(\mu) - V_n(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) \quad (4.12)$$

for all  $n \in \mathbb{N}$ . The property is obviously satisfied for  $n = 0$ . Assume that the property holds true for a fixed  $n \in \mathbb{N}$ , and let us prove it for  $n + 1$ . First of all, the inequality (4.2) implies that  $V_n$  is continuous, and thus  $V_n \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ . Therefore,  $V_{n+1} = \mathcal{T}V_n$  is well defined. Fix  $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ . In order to use the expression (4.1) of the Bellman operator  $\mathcal{T}$ , we consider an optimal coupling  $(\xi, \xi')$  of  $\mu$  and  $\mu'$ , i.e.  $\xi \sim \mu$ ,  $\xi' \sim \mu'$ , and  $\mathbb{E}[d(\xi, \xi')] = \mathcal{W}(\mu, \mu')$ , and fix an  $A$ -valued random variable  $\alpha_0$ . Let us start with two preliminary estimations: under  $(\mathbf{H}_{\text{ip}})$ , we have

$$\begin{aligned}\mathbb{E}[|f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)) - f(\xi', \alpha_0, \mathcal{L}(\xi', \alpha_0))|] &\leq K_f(\mathbb{E}[d(\xi, \xi')] + \mathcal{W}(\mathcal{L}(\xi, \alpha_0), \mathcal{L}(\xi', \alpha_0))) \\ &\leq K_f(\mathbb{E}[d(\xi, \xi')] + \mathbb{E}[d((\xi, \alpha_0), (\xi', \alpha_0))]) \\ &\leq 2K_f \mathbb{E}[d(\xi, \xi')] = 2K_f \mathcal{W}(\mu, \mu').\end{aligned} \quad (4.13)$$

Similarly, for  $e^0 \in E^0$ , we have

$$\mathbb{E}[d(F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, e^0), F(\xi', \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, e^0))] \leq 2K_F \mathcal{W}(\mu, \mu'). \quad (4.14)$$

Now, we prove the hereditary property. The definition of  $\mathcal{T}$  and  $V_{n+1}$  combined with (4.2) and the induction hypothesis, imply that

$$|V_{n+1}(\mu) - V_{n+1}(\mu')| \leq 2K_f \mathcal{W}(\mu, \mu') + \beta \mathbb{E}[2K_f \sum \beta^t \min((2K_F^t \mathcal{W}(\mu_1, \mu'_1), \Delta_{\mathcal{X}})]$$

where  $\mu_1 = \mathcal{L}^0(F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, \varepsilon_1^0))$  and  $\mu'_1 = \mathcal{L}^0(F(\xi', \alpha_0, \mathcal{L}(\xi', \alpha_0), \varepsilon_1^1, \varepsilon_1^0))$ . By Jensen's inequality and (4.2), we have

$$\begin{aligned} |V_{n+1}(\mu) - V_{n+1}(\mu')| &\leq 2K_f \min(\mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t \mathbb{E} \mathcal{W}(\mu_1, \mu'_1), \Delta_{\mathcal{X}}) \\ &\leq 2K_f \min(\mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t 2K_F \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) \\ &\leq 2K_f \sum \beta^t \min((2K_F^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}). \end{aligned}$$

This concludes the induction and proves that  $V_n$  is well defined and satisfies the inequality (4.2) for all  $n \in \mathbb{N}$ . As  $\mathcal{T}$  is  $\beta$ -contracting, a standard argument from the proof of the Banach fixed point theorem shows that  $(V_n)_n$  is a Cauchy sequence in the complete metric space  $L_m^\infty(\mathcal{P}(\mathcal{X}))$ , and therefore admits a limit  $V^* \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ . Notice that

$$V^*(\mu) = \lim_n V_{n+1}(\mu) = \lim_n \mathcal{T}V_n(\mu) = \mathcal{T}V^*$$

by continuity of the contracting operator  $\mathcal{T}$ .

(iii) By sending  $n$  to infinity in (4.2), we obtain

$$|V(\mu) - V(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}).$$

Then, a similar derivation as in the end of proof of Theorem 2.1 shows the required  $\gamma$ -Hölder property of  $V^*$ .  $\square$

**Remark 4.2** In the proof of Proposition 4.2, one could also have proved that the set  $\mathcal{S}$  of functions  $W : \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$  such that

$$|W(\mu) - W(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}})$$

for all  $\mu, \mu' \in \mathcal{P}(\mathcal{X})$  is a complete metric space, as it is a closed set of the complete metric space  $L_m^\infty(\mathcal{P}(\mathcal{X}))$ , and is stabilized by the contracting operator  $\mathcal{T}$  (which is essentially proved by replacing  $V_n$  by  $W$  in the proof). One could then have invoked the Banach fixed point theorem on this set  $\mathcal{S}$ , implying the existence and uniqueness of the fixed point  $V^*$ . Notice that this argument would not work if we considered, instead of  $\mathcal{S}$ , the set of  $\gamma$ -Hölder continuous functions. Indeed, while it is true that such set is stabilized by  $\mathcal{T}$  (it essentially follows from (4.2) and (4.2)), the set of  $\gamma$ -Hölder continuous functions is not closed in  $L_m^\infty(\mathcal{P}(\mathcal{X}))$  (and thus not a complete metric space): there might indeed exist a converging sequence of  $\gamma$ -Hölder continuous functions with multiplicative factors (in the Hölder property) tending toward infinity, such that the limit function is not  $\gamma$ -Hölder anymore.  $\square$

As a consequence of Proposition 4.2, we can easily show the following relation between the value function  $\tilde{V}$  of the general lifted MDP, and the fixed point  $V^*$  of the Bellman operator.



**Lemma 4.3** For all  $\mu \in \mathcal{P}(\mathcal{X})$ , we have  $\tilde{V}(\mu) \leq V^*(\mu)$ .

**Proof.** From (4.2), we have

$$\begin{aligned} & \inf_{\mu \in \mathcal{P}(\mathcal{X})} \{V^*(\mu) - \tilde{V}^\nu(\mu)\} \\ & \geq \inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \mathcal{T}V^*(\mu) - \left( \tilde{f}(\mu, \nu_0) + \beta \mathbb{E} \left[ V^*(\mu_1^{\mu, \nu}) \right] \right) + \beta \mathbb{E} \left[ V^*(\mu_1^{\mu, \nu}) - \tilde{V}^{\tilde{\nu}^{\varepsilon_1^0}}(\mu_1^{\mu, \nu}) \right] \right\} \\ & \geq \beta \mathbb{E} \left[ V^*(\mu_1^{\mu, \nu}) - \tilde{V}^{\tilde{\nu}^{\varepsilon_1^0}}(\mu_1^{\mu, \nu}) \right] \geq \beta \inf_{\mu \in \mathcal{P}(\mathcal{X})} \{V^*(\mu) - \tilde{V}^\nu(\mu)\}. \end{aligned}$$

This shows that  $\inf_{\mu \in \mathcal{P}(\mathcal{X})} (V^*(\mu) - \tilde{V}^\nu(\mu)) \geq 0$ , hence

$$\tilde{V}^\nu(\mu) \leq V^*(\mu) \quad \forall \mu \in \mathcal{P}(\mathcal{X}).$$

Taking the sup over  $\nu \in \mathcal{A}$ , we obtain the required result.  $\square$

### 4.3 Building $\epsilon$ -optimal randomized feedback controls

We aim to prove rigorously the equality  $\tilde{V} = V^*$ , i.e., the value function  $\tilde{V}$  of the general lifted MDP satisfies the Bellman fixed point equation:  $\tilde{V} = \mathcal{T}\tilde{V}$ , and also to show the existence of an  $\epsilon$ -optimal control for  $\tilde{V}$ . Notice that it cannot be obtained directly from classical theory of MDP as we consider here open-loop controls  $\nu \in \mathcal{A}$  while MDP usually deals with feedback controls on finite-dimensional spaces. Anyway, following the standard notation in MDP theory with state space  $\mathcal{P}(\mathcal{X})$  and action space  $\mathbf{A}$ , and in connection with the Bellman operator in (4.2), we introduce, for  $\pi \in L^0(\mathcal{P}(\mathcal{X}); \mathbf{A})$  (the set of measurable functions from  $\mathcal{P}(\mathcal{X})$  into  $\mathbf{A}$ ) called (measurable) feedback policy, the so-called  $\pi$ -Bellman operator  $\mathcal{T}^\pi$  on  $L^\infty(\mathcal{P}(\mathcal{X}))$ , defined for  $W \in L^\infty(\mathcal{P}(\mathcal{X}))$  by

$$[\mathcal{T}^\pi W](\mu) = \tilde{f}(\mu, \pi(\mu)) + \beta \mathbb{E} [W(\tilde{F}(\mu, \pi(\mu), \varepsilon_1^0))], \quad \mu \in \mathcal{P}(\mathcal{X}). \quad (4.15)$$

As for the Bellman operator  $\mathcal{T}$ , we have the basic properties on the operator  $\mathcal{T}^\pi$ .

**Lemma 4.4** Fix  $\pi \in L^0(\mathcal{P}(\mathcal{X}); \mathbf{A})$ .

- (i) The operator  $\mathcal{T}^\pi$  is contracting on  $L^\infty(\mathcal{P}(\mathcal{X}))$  with Lipschitz factor  $\beta$ , and admits a unique fixed point denoted  $\tilde{V}^\pi$ .
- (ii) Furthermore, it is monotone increasing: for  $W_1, W_2 \in L^\infty(\mathcal{P}(\mathcal{X}))$ , if  $W_1 \leq W_2$ , then  $\mathcal{T}^\pi W_1 \leq \mathcal{T}^\pi W_2$ .

**Remark 4.3** It is well-known from MDP theory that the fixed point  $\tilde{V}^\pi$  to the operator  $\mathcal{T}^\pi$  is equal to

$$\tilde{V}^\pi(\mu) = \mathbb{E} \left[ \sum_{t \in \mathbb{N}} \tilde{f}(\mu_t, \pi(\mu_t)) \right],$$

where  $(\mu_t)$  is the MDP in (4.1) with the feedback and stationary control  $\nu^\pi = (\nu_t^\pi)_t \in \mathcal{A}$  defined by  $\nu_t^\pi = \pi(\mu_t)$ ,  $t \in \mathbb{N}$ . In the sequel, we shall then identify by misuse of notation  $\tilde{V}^\pi$  and  $\tilde{V}^{\nu^\pi}$  as defined in (4.1).  $\square$

Our ultimate goal being to solve the CMKV-MDP, we introduce a subclass of feedback policies for the lifted MDP.

**Definition 4.1 (Lifted randomized feedback policy)**

A feedback policy  $\pi \in L^0(\mathcal{P}(\mathcal{X}); \mathbf{A})$  is a lifted randomized feedback policy if there exists a measurable function  $\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1]; A)$ , called randomized feedback policy, such that  $(\xi, \mathbf{a}(\mu, \xi, U)) \sim \pi(\mu)$ , for all  $\mu \in \mathcal{P}(\mathcal{X})$ , with  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ .

**Remark 4.4** Given  $\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1]; A)$ , denote by  $\pi^\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}); \mathbf{A})$  the associated lifted randomized feedback policy, i.e.,  $\pi^\mathbf{a}(\mu) = \mathcal{L}(\xi, \mathbf{a}(\mu, \xi, U))$ , for  $\mu \in \mathcal{P}(\mathcal{X})$ , and  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ . By definition of the  $\pi$ -Bellman operator  $\mathcal{T}^{\pi^\mathbf{a}}$  in (4.3), and observing that  $\mathbf{p}(\mu, \pi^\mathbf{a}(\mu)) = \pi^\mathbf{a}(\mu) = \mathcal{L}(\xi, \mathbf{a}^\mu(\xi, U))$ , where we set  $\mathbf{a}^\mu = \mathbf{a}(\mu, \cdot, \cdot) \in L^0(\mathcal{X} \times [0, 1]; A)$ , we see (recalling the notation in (4.1)) that for all  $W \in L^\infty(\mathcal{P}(\mathcal{X}))$ ,

$$[\mathcal{T}^{\pi^\mathbf{a}} W](\mu) = [\mathbb{T}^{\mathbf{a}^\mu} W](\mu), \quad \mu \in \mathcal{P}(\mathcal{X}). \quad (4.16)$$

On the other hand, let  $\xi \in L^0(\mathcal{G}; \mathcal{X})$  be some initial state satisfying the randomization hypothesis **Rand**( $\xi, \mathcal{G}$ ), and denote by  $\alpha^\mathbf{a} \in \mathcal{A}$  the randomized feedback stationary control defined by  $\alpha_t^\mathbf{a} = \mathbf{a}(\mathbb{P}_{X_t}^0, X_t, U_t)$ , where  $X = X^{\xi, \alpha^\mathbf{a}}$  is the state process in (3) of the CMKV-MDP, and  $(U_t)_t$  is an i.i.d. sequence of uniform  $\mathcal{G}$ -measurable random variables independent of  $\xi$ . By construction, the associated lifted control  $\alpha^\mathbf{a} = \mathcal{L}_\xi^0(\alpha^\mathbf{a})$  satisfies  $\alpha_t^\mathbf{a} = \mathbb{P}_{(X_t, \alpha_t^\mathbf{a})}^0 = \pi^\mathbf{a}(\mu_t)$ , where  $\mu_t = \mathbb{P}_{X_t}^0$ ,  $t \in \mathbb{N}$ . Denoting by  $V^\mathbf{a} := V^{\alpha^\mathbf{a}}$  the associated expected gain of the CMKV-MDP, and recalling Remark 4.3, we see from (4.1) that  $V^\mathbf{a}(\xi) = \tilde{V}^{\nu^{\pi^\mathbf{a}}}(\mu) = \tilde{V}^{\pi^\mathbf{a}}(\mu)$ , where  $\mu = \mathcal{L}(\xi)$ .  $\square$

We show a verification type result for the general lifted MDP, and as a byproduct for the CMKV-MDP, by means of the Bellman operator.

**Proposition 4.3 (Verification result)**

Fix  $\epsilon \geq 0$ , and suppose that there exists an  $\epsilon$ -optimal feedback policy  $\pi_\epsilon \in L^0(\mathcal{P}(\mathcal{X}); \mathbf{A})$  for  $V^*$  in the sense that

$$V^* \leq \mathcal{T}^{\pi_\epsilon} V^* + \epsilon.$$

Then,  $\nu^{\pi_\epsilon} \in \mathcal{A}$  is  $\frac{\epsilon}{1-\beta}$ -optimal for  $\tilde{V}$ , i.e.,  $\tilde{V}^{\pi_\epsilon} \geq \tilde{V} - \frac{\epsilon}{1-\beta}$ , and we have  $\tilde{V} \geq V^* - \frac{\epsilon}{1-\beta}$ . Furthermore, if  $\pi_\epsilon$  is a lifted randomized feedback policy, i.e.,  $\pi_\epsilon = \pi^{\mathbf{a}_\epsilon}$ , for some  $\mathbf{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1]; A)$ , then under **Rand**( $\xi, \mathcal{G}$ ),  $\alpha^{\mathbf{a}_\epsilon} \in \mathcal{A}$  is an  $\frac{\epsilon}{1-\beta}$ -optimal control for  $V(\xi)$ , i.e.,  $V^{\mathbf{a}_\epsilon}(\xi) \geq V(\xi) - \frac{\epsilon}{1-\beta}$ , and we have  $V(\xi) \geq V^*(\mu) - \frac{\epsilon}{1-\beta}$ , for  $\mu = \mathcal{L}(\xi)$ .

**Proof.** Since  $\tilde{V}^{\pi_\epsilon} = \mathcal{T}^{\pi_\epsilon} \tilde{V}^{\pi_\epsilon}$ , and recalling from Lemma 4.3 that  $V^* \geq \tilde{V} \geq \tilde{V}^{\pi_\epsilon}$ , we have for all  $\mu \in \mathcal{P}(\mathcal{X})$ ,

$$\left| (V^* - \tilde{V}^{\pi_\epsilon})(\mu) \right| \leq \left| \mathcal{T}^{\pi_\epsilon} (V^* - \tilde{V}^{\pi_\epsilon})(\mu) + \epsilon \right| \leq \beta \|V^* - \tilde{V}^{\pi_\epsilon}\| + \epsilon,$$

where we used the  $\beta$ -contraction property of  $\mathcal{T}^{\pi_\epsilon}$  in Lemma 4.4. We deduce that  $\|V^* - \tilde{V}^{\pi_\epsilon}\| \leq \frac{\epsilon}{1-\beta}$ , and then,  $\tilde{V} \geq \tilde{V}^{\pi_\epsilon} \geq V^* - \frac{\epsilon}{1-\beta}$ , which combined with  $V^* \geq \tilde{V}$ , shows the first assertion. Moreover, if  $\pi_\epsilon = \pi^{\mathbf{a}_\epsilon}$  is a lifted randomized feedback policy, then by Remark 4.4, and under **Rand**( $\xi, \mathcal{G}$ ), we have  $V^{\mathbf{a}_\epsilon}(\xi) = \tilde{V}^{\pi_\epsilon}(\mu)$ . Recalling that  $V(\xi) \leq \tilde{V}(\mu)$ , and together with the first assertion, this proves the required result.  $\square$

**Remark 4.5** If we can find for any  $\epsilon > 0$ , an  $\epsilon$ -optimal lifted randomized feedback policy for  $V^*$ , then according to Proposition 4.3, and under **Rand**( $\xi, \mathcal{G}$ ), one could restrict to randomized feedback policies in the computation of the optimal value  $V(\xi)$  of the CMKV-MDP, i.e.,  $V(\xi) = \sup_{\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)} V^{\mathbf{a}}(\xi)$ . Moreover, this would prove that  $V(\xi) = \tilde{V}(\mu) = V^*(\mu)$ , hence  $V$  is law-invariant, and satisfies the Bellman fixed equation.

Notice that instead of proving directly the dynamic programming Bellman equation for  $V$ , we start from the fixed point solution  $V^*$  to the Bellman equation, and show via a verification result that  $V$  is indeed equal to  $V^*$ , hence satisfies the Bellman equation.

By the formulation (4.1) of the Bellman operator in Proposition 4.1, and the fixed point equation satisfied by  $V^*$ , we know that for all  $\epsilon > 0$ , and  $\mu \in \mathcal{P}(\mathcal{X})$ , there exists  $\mathbf{a}_\epsilon^\mu \in L^0(\mathcal{X} \times [0,1]; A)$  such that

$$V^*(\mu) \leq [\mathbb{T}^{\mathbf{a}_\epsilon^\mu} V^*](\mu) + \epsilon. \quad (4.17)$$

The crucial issue is to prove that the mapping  $(\mu, x, u) \mapsto \mathbf{a}_\epsilon(\mu, x, u) := \mathbf{a}_\epsilon^\mu(x, u)$  is measurable so that it defines a randomized feedback policy  $\mathbf{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$ , and an associated lifted randomized feedback policy  $\pi^{\mathbf{a}_\epsilon}$ . Recalling the relation (4.4), this would then show that  $\pi^{\mathbf{a}_\epsilon}$  is a  $\epsilon$ -optimal lifted randomized feedback policy for  $V^*$ , and we could apply the verification result.  $\square$

We now address the measurability issue for proving the existence of an  $\epsilon$ -optimal randomized feedback policy for  $V^*$ . The basic idea is to construct as in (4.5) an  $\epsilon$ -optimal  $\mathbf{a}_\epsilon^\mu \in L^0(\mathcal{X} \times [0,1]; A)$  for  $V^*(\mu)$  when  $\mu$  lies in a suitable finite grid of  $\mathcal{P}(\mathcal{X})$ , and then “patches” things together to obtain an  $\epsilon$ -optimal randomized feedback policy. This is made possible under some uniform continuity property of  $V^*$ .

The next result provides a suitable discretization of the set of probability measures.

**Lemma 4.5 (Quantization of  $\mathcal{P}(\mathcal{X})$ )**

Fix  $\eta > 0$ . Then for each finite  $\eta/2$ -covering  $\mathcal{X}_\eta$  of  $\mathcal{X}$ , one can construct a finite subset  $\mathcal{M}_\eta$  of  $\mathcal{P}(\mathcal{X})$ , of size  $N_\eta = n_\eta^{\#\mathcal{X}_\eta - 1}$ , where  $n_\eta$  is a grid size of  $[0,1]$ , that is an  $\eta$ -covering of  $\mathcal{P}(\mathcal{X})$ .

**Proof.** As  $\mathcal{X}$  is compact, there exists a finite subset  $\mathcal{X}_\eta \subset \mathcal{X}$  such that  $d(x, x_\eta) \leq \eta/2$  for all  $x \in \mathcal{X}$ , where  $x_\eta$  denotes the projection of  $x$  on  $\mathcal{X}_\eta$ . Given  $\mu \in \mathcal{P}(\mathcal{X})$ , and  $\xi \sim \mu$ , we denote by  $\xi_\eta$  the quantization, i.e., the projection of  $\xi$  on  $\mathcal{X}_\eta$ , and by  $\mu_\eta$  the discrete law of  $\xi_\eta$ . Thus,  $\mathbb{E}[d(\xi, \xi_\eta)] \leq \eta/2$ , and therefore  $\mathcal{W}(\mu, \mu_\eta) \leq \eta/2$ . The probability measure  $\mu_\eta$  lies in  $\mathcal{P}(\mathcal{X}_\eta)$ , which is identified with the simplex of  $[0,1]^{\#\mathcal{X}_\eta}$ . We then use another grid  $G_\eta = \{\frac{i}{n_\eta} : i = 0, \dots, n_\eta\}$  of  $[0,1]$ , and project its weights  $\mu_\eta(y) \in [0,1]$ ,  $y \in \mathcal{X}_\eta$ , on  $G_\eta$ , in order to obtain another discrete probability measure  $\mu_{\eta, n_\eta}$ . From the dual Kantorovich representation of Wasserstein distance, it is easy to see that for  $n_\eta$  large enough,  $\mathcal{W}(\mu_\eta, \mu_{\eta, n_\eta}) \leq \eta/2$ , and so  $\mathcal{W}(\mu, \mu_{\eta, n_\eta}) \leq \eta$ . We conclude the proof by noting that  $\mu_{\eta, n_\eta}$  belongs to the set  $\mathcal{M}_\eta$  of probability measures on  $\mathcal{X}_\eta$  with weights valued in the finite grid  $G_\eta$ , hence  $\mathcal{M}_\eta$  is a finite set of  $\mathcal{P}(\mathcal{X}_\eta)$ , of cardinal  $N_\eta = n_\eta^{\#\mathcal{X}_\eta - 1}$ .  $\square$

**Remark 4.6** Lemma 4.5 is actually a simple consequence of Prokhorov’s theorem, but the above proof has some advantages:

- it provides an explicit construction of the quantization grid  $\mathcal{M}_\eta$ ,
- it explicitly gives the size of the grid as a function of  $\eta$ , which is particularly useful for computing the time/space complexity of algorithms,

- this special grid simultaneously quantizes  $\mathcal{P}(\mathcal{X})$  and  $\mathcal{X}$ , in the sense that the measures from  $\mathcal{M}_\eta$  are all supported on the finite set  $\mathcal{X}_\eta$ . This is also useful for algorithms because for  $\mu \in \mathcal{M}_\eta$ ,  $\hat{a} \in \hat{A}(\mathcal{X})$ , and  $W \in L^\infty(\mathbb{R})$ , the expression  $\hat{T}^{\hat{a}}W(\mu)$  only depends upon  $(\hat{a}(x))_{x \in \mathcal{X}_\eta}$ . Therefore, in the Bellman fixed point equation, the computation of an  $\epsilon$ -argmax over the set  $L^0(\mathcal{X}, \mathcal{P}(A))$  is reduced to a computation of an  $\epsilon$ -argmax over the set  $\mathcal{P}(A)^{\mathcal{X}_\eta}$ , which is more tractable for a computer.

We can conclude this paragraph by showing the existence of an  $\epsilon$ -optimal lifted randomized feedback policy for the general lifted MDP on  $\mathcal{P}(\mathcal{X})$ , and obtain as a by-product the corresponding Bellman fixed point equation for its value function and for the optimal value of the CMKV-MDP under randomization hypothesis.

**Theorem 4.1** *Assume that  $(\mathbf{H}_{\text{lip}})$  holds true. Then, for all  $\epsilon > 0$ , there exists a lifted randomized feedback policy  $\pi^{\mathbf{a}_\epsilon}$ , for some  $\mathbf{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1]; A)$ , that is  $\epsilon$ -optimal for  $V^*$ . Consequently, under  $\mathbf{Rand}(\xi, \mathcal{G})$ , the randomized feedback stationary control  $\alpha^{\mathbf{a}_\epsilon} \in \mathcal{A}$  is  $\frac{\epsilon}{1-\beta}$ -optimal for  $V(\xi)$ , and we have  $V(\xi) = \tilde{V}(\mu) = V^*(\mu)$ , for  $\mu = \mathcal{L}(\xi)$ , which thus satisfies the Bellman fixed point equation.*

**Proof.** Fix  $\epsilon > 0$ , and given  $\eta > 0$ , consider a quantizing grid  $\mathcal{M}_\eta = \{\mu^1, \dots, \mu^{N_\eta}\} \subset \mathcal{P}(\mathcal{X})$  as in Lemma 4.5, and an associated partition  $C_\eta^i$ ,  $i = 1, \dots, N_\eta$ , of  $\mathcal{P}(\mathcal{X})$ , satisfying

$$C_\eta^i \subset B_\eta(\mu^i) := \left\{ \mu \in \mathcal{P}(\mathcal{X}) : \mathcal{W}(\mu, \mu^i) \leq \eta \right\}, \quad i = 1, \dots, N_\eta.$$

For any  $\mu^i$ ,  $i = 1, \dots, N_\eta$ , and by (4.5), there exists  $\mathbf{a}_\epsilon^i \in L^0(\mathcal{X} \times [0, 1]; A)$  such that

$$V^*(\mu^i) \leq [\mathbb{T}^{\mathbf{a}_\epsilon^i} V^*](\mu^i) + \frac{\epsilon}{3}. \quad (4.18)$$

From the partition  $C_\eta^i$ ,  $i = 1, \dots, N_\eta$  of  $\mathcal{P}(\mathcal{X})$ , associated to  $\mathcal{M}_\eta$ , we construct the function  $\mathbf{a}_\epsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1] \rightarrow A$  as follows. Let  $h_1, h_2$  be two measurable functions from  $[0, 1]$  into  $[0, 1]$ , such that if  $U \sim \mathcal{U}([0, 1])$ , then  $(h_1(U), h_2(U)) \sim \mathcal{U}([0, 1])^{\otimes 2}$ . We then define

$$\mathbf{a}_\epsilon(\mu, x, u) = \mathbf{a}_\epsilon^i(\zeta(\mu, \mu^i, x, h_1(u)), h_2(u)), \quad \text{when } \mu \in C_\eta^i, \quad i = 1, \dots, N_\eta, \quad x \in \mathcal{X}, \quad u \in [0, 1],$$

where  $\zeta$  is the measurable coupling function defined in Lemma 4.1. Such function  $\mathbf{a}_\epsilon$  is clearly measurable, i.e.,  $\mathbf{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1]; A)$ , and we denote by  $\pi_\epsilon = \pi^{\mathbf{a}_\epsilon}$  the associated lifted randomized feedback policy, which satisfies

$$[\mathcal{T}^{\pi_\epsilon} V^*](\mu^i) = [\mathbb{T}^{\mathbf{a}_\epsilon^i} V^*](\mu^i), \quad i = 1, \dots, N_\eta, \quad (4.19)$$

by (4.4). Let us now check that such  $\pi_\epsilon$  yields an  $\epsilon$ -optimal randomized feedback policy for  $\eta$  small enough. For  $\mu \in \mathcal{P}(\mathcal{X})$ , with  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ , we set  $U_1 := h_1(U)$ ,  $U_2 := h_2(U)$ , and define  $\mu_\eta = \mu^i$ , when  $\mu \in C_\eta^i$ ,  $i = 1, \dots, N_\eta$ , and  $\xi_\eta := \zeta(\mu, \mu_\eta, \xi, U_1)$ . Observe by Lemma 4.5 that  $\mathcal{W}(\mu, \mu_\eta) \leq \eta$ , and by Lemma 4.1 that  $(\xi_\eta, U_2) \sim \mu_\eta \otimes \mathcal{U}([0, 1])$ . We then write for any  $\mu \in \mathcal{P}(\mathcal{X})$ ,

$$\begin{aligned} [\mathcal{T}^{\pi_\epsilon} V^*](\mu) - V^*(\mu) &= \left( [\mathcal{T}^{\pi_\epsilon} V^*](\mu) - [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta) \right) + \left( [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta) - V^*(\mu_\eta) \right) \\ &\quad + (V^*(\mu_\eta) - V^*(\mu)) \\ &\geq \left( [\mathcal{T}^{\pi_\epsilon} V^*](\mu) - [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta) \right) - \frac{\epsilon}{3} - \frac{\epsilon}{3}, \end{aligned} \quad (4.20)$$

where we used (4.3)-(4.3) and the fact that  $|V^*(\mu_\eta) - V^*(\mu)| \leq \epsilon/3$  for  $\eta$  small enough by uniform continuity of  $V^*$  in Proposition 4.2. Moreover, by observing that  $\mathbf{a}_\epsilon(\mu, \xi, U) = \mathbf{a}_\epsilon(\mu_\eta, \xi_\eta, U_2) =: \alpha_0$ , so that  $\pi_\epsilon(\mu) = \mathcal{L}(\xi, \alpha_0)$ ,  $\pi_\epsilon(\mu_\eta) = \mathcal{L}(\xi_\eta, \alpha_0)$ , we have

$$\begin{aligned} [\mathcal{T}^{\pi_\epsilon} V^*](\mu) &= \mathbb{E} \left[ f(Y) + \beta V^*(\mathbb{P}_{F(Y, \varepsilon_1, \varepsilon_1^0)}^0) \right], \\ [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta) &= \mathbb{E} \left[ f(Y_\eta) + \beta V^*(\mathbb{P}_{F(Y_\eta, \varepsilon_1, \varepsilon_1^0)}^0) \right], \end{aligned}$$

where  $Y = (\xi, \alpha_0, \pi_\epsilon(\mu))$ , and  $Y_\eta = (\xi_\eta, \alpha_0, \pi_\epsilon(\mu_\eta))$ . Under  $(\mathbf{H}_{\text{lip}})$ , by using the  $\gamma$ -Hölder property of  $V^*$  with constant  $K_*$  in Proposition 4.2, and by definition of the Wasserstein distance (recall that  $\xi \sim \mu$ ,  $\xi_\eta \sim \mu_\eta$ ), we then get

$$\begin{aligned} |[\mathcal{T}^{\pi_\epsilon} V^*](\mu) - [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta)| &\leq 2K \mathbb{E} [d(\xi, \xi_\eta)] \\ &\quad + \beta K_* \mathbb{E} \left[ \mathbb{E} [d(F(\xi, \alpha_0, \pi_\epsilon(\mu), \varepsilon_1, e), F(\xi_\eta, \alpha_0, \pi_\epsilon(\mu_\eta), \varepsilon_1, e))]^\gamma]_{e:=\varepsilon_1^0} \right] \\ &\leq 2K \mathbb{E} [d(\xi, \xi_\eta)] \\ &\quad + \beta K_* \mathbb{E} \left[ \mathbb{E} [d(F(\xi, \alpha_0, \pi_\epsilon(\mu), \varepsilon_1, e), F(\xi_\eta, \alpha_0, \pi_\epsilon(\mu_\eta), \varepsilon_1, e))]^\gamma]_{e:=\varepsilon_1^0} \right]^\gamma \\ &\leq C \mathbb{E} [d(\xi, \xi_\eta)]^\gamma. \end{aligned}$$

for some constant  $C$  independent from  $\mathbb{E} [d(\xi, \xi_\eta)]$ . Now, by the coupling Lemma 4.1, one can choose  $\eta$  small enough so that  $C \mathbb{E} [d(\xi, \xi_\eta)]^\gamma \leq \frac{\epsilon}{3}$ . Therefore,  $|[\mathcal{T}^{\pi_\epsilon} V^*](\mu) - [\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta)| \leq \epsilon/3$ , and, plugging into (4.3), we obtain  $\mathcal{T}^{\pi_\epsilon} V^*(\mu) - V^*(\mu) \geq -\epsilon$ , for all  $\mu \in \mathcal{P}(\mathcal{X})$ , which means that  $\pi_\epsilon$  is  $\epsilon$ -optimal for  $V^*$ . The rest of the assertions in the Theorem follows from the verification result in Proposition 4.3.  $\square$

**Remark 4.7** We stress the importance of the coupling Lemma in the construction of  $\epsilon$ -optimal control in Theorem 4.1. Indeed, as we do not make any regularity assumption on  $F$  and  $f$  with respect to the “control arguments”, the only way to make  $[\mathcal{T}^{\pi_\epsilon} V^*](\mu)$  and  $[\mathcal{T}^{\pi_\epsilon} V^*](\mu_\eta)$  close to each other is to couple terms to have the same control in  $F$  and  $f$ . This is achieved by turning  $\mu$  into  $\mu_\eta$ ,  $\xi$  into  $\xi_\eta$  and set  $\alpha_0 = \mathbf{a}_\epsilon(\mu, \xi, U) = \mathbf{a}_\epsilon(\mu_\eta, \xi_\eta, U_2)$ . Turning  $\mu$  into  $\mu_\eta$  is a simple quantization, but turning  $\xi$  into  $\xi_\eta$  is obtained thanks to the coupling Lemma.  $\square$

**Remark 4.8** Theorem 4.1, although applying to a more general case than the results from Section 3, provides a weaker result. Indeed, it does not state that any control  $\nu$  for the lifted MDP  $\tilde{V}(\mu)$  can be represented, i.e., associated to a control  $\alpha$  for  $V(\xi)$  such that  $\alpha_t := \mathbb{P}_{(X_t, \alpha_t)}^0 = \nu_t$  for all  $t \in \mathbb{N}$ . This theorem only implies that one can *restrict* the optimization to representable controls without changing the optimal value. Consequently, contrarily to Theorem 3.1 and Theorem 3.2, here one cannot conclude that an optimal control for  $V(\xi)$  exists iff an optimal control for  $\tilde{V}(\mu)$  exists. More precisely, it is possible that an optimal control  $\nu$  for  $\tilde{V}(\mu)$  exists but cannot be associated to a control  $\alpha$  for  $V(\xi)$  such that  $\alpha = \nu$ , and thus the existence of an optimal control for  $\tilde{V}(\mu)$  does not guarantee the existence of an optimal control for  $V(\xi)$ .  $\square$

#### 4.4 Relaxing the randomization hypothesis

We now relax the randomization hypothesis, and make the weaker assumption (recall Remark 3.1) that the initial information filtration  $\mathcal{G}$  is rich enough.

We need to state some uniform continuity property on the value function  $V$  of the CMKV-MDP.

**Lemma 4.6** *Assume that  $(\mathbf{H}_{\text{lip}})$  holds true. Then, there exists a constant  $K$  (only depending upon  $K_F, \beta$ , and  $\Delta_{\mathcal{X}}$ ) such that*

$$\sup_{\alpha \in \mathcal{A}} |V^\alpha(\xi) - V^\alpha(\xi')| \leq K_\gamma (\mathbb{E}[d(\xi, \xi')])^\gamma, \quad \forall \xi, \xi' \in L^0(\mathcal{G}; \mathcal{X}),$$

where  $\gamma = \min(1, \frac{|\ln(\beta)|}{(\ln 2K)_+})$ . Consequently,  $V$  is  $\gamma$ -Hölder on  $L^0(\mathcal{G}; \mathcal{X})$  endowed with the  $L^1$ -distance.

**Proof.** Fix  $\xi, \xi' \in L^0(\Omega, \mathcal{X})$ , and consider an arbitrary  $\alpha = \alpha^\pi \in \mathcal{A}$  associated to an open-loop policy  $\pi \in \Pi_{OL}$ . By Proposition A.1, there exists a measurable function  $f_{t,\pi} \in L^0(\mathcal{X} \times G \times E^t \times (E^0)^t; \mathcal{X})$ , s.t.  $X_t^{\xi, \alpha} = f_{t,\pi}(\xi, \Gamma, (\varepsilon_s)_{s \leq t}, (e_s^0)_{s \leq t})$ , thus  $\mathbb{P}_{X_t^{\xi, \alpha}}^0 = \mathcal{L}(f_{t,\pi}(\xi, \Gamma, (\varepsilon_s)_{s \leq t}, (e_s^0)_{s \leq t}))_{e_s^0 = \varepsilon_s^0, s \leq t}$ . We thus have

$$\mathcal{W}(\mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \mathbb{P}_{(X_t^{\xi', \alpha}, \alpha_t)}^0) \leq \mathbb{E} \left[ d \left( f_{t,\pi}(\xi, \Gamma, (\varepsilon_s)_{s \leq t}, (e_s^0)_{s \leq t}), f_{t,\pi}(\xi', \Gamma, (\varepsilon_s)_{s \leq t}, (e_s^0)_{s \leq t}) \right) \right]_{e_s^0 = \varepsilon_s^0, s \leq t},$$

and so

$$\mathbb{E} \left[ \mathcal{W}(\mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \mathbb{P}_{(X_t^{\xi', \alpha}, \alpha_t)}^0) \right] \leq \mathbb{E} [d(X_t^{\xi, \alpha}, X_t^{\xi', \alpha})], \quad (4.21)$$

Under the Lipschitz condition on  $f$  in  $(\mathbf{H}_{\text{lip}})$ , we then have

$$|V^\alpha(\xi) - V^\alpha(\xi')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \mathbb{E} [d(X_t^{\xi, \alpha}, X_t^{\xi', \alpha})]. \quad (4.22)$$

By conditioning, and from the transition dynamics of the state process, we see that for  $t \in \mathbb{N}$ ,  $\mathbb{E} [d(X_{t+1}^{\xi, \alpha}, X_{t+1}^{\xi', \alpha})] = \mathbb{E} [\Delta(\alpha_t^i, X_t^{\xi, \alpha}, \mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, X_t^{\xi', \alpha}, \mathbb{P}_{(X_t^{\xi', \alpha}, \alpha_t)}^0, \varepsilon_{t+1}^0)]$ , where

$$\Delta(a, x, \nu, x', \nu', e_{t+1}^0) = \mathbb{E} [d(F(x, a, \nu, \varepsilon_{t+1}, e_{t+1}^0), F(x', a, \nu', \varepsilon_{t+1}, e_{t+1}^0))].$$

By the Lipschitz condition on  $F$  in  $(\mathbf{H}_{\text{lip}})$ , we thus have

$$\mathbb{E} [d(X_{t+1}^{\xi, \alpha}, X_{t+1}^{\xi', \alpha})] \leq K_F \mathbb{E} [d(X_t^{\xi, \alpha}, X_t^{\xi', \alpha}) + \mathcal{W}(\mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \mathbb{P}_{(X_t^{\xi', \alpha}, \alpha_t)}^0)] \leq 2K_F \mathbb{E} [d(X_t^{\xi, \alpha}, X_t^{\xi', \alpha})],$$

where the last inequality comes from (4.4). Denoting by  $\delta_t(\xi, \xi') := \sup_{\alpha \in \mathcal{A}} \mathbb{E} [d(X_t^{\xi, \alpha}, X_t^{\xi', \alpha})]$ , and noting that  $\delta_0(\xi, \xi') = \mathbb{E} d(\xi, \xi')$ , it follows by induction that there exists  $K \in \mathbb{R}$  such that

$$\delta_t(\xi, \xi') \leq s_t(\mathbb{E} d(\xi, \xi')), \quad s_t(m) := m(2K)^t, \quad m \geq 0, \quad t \in \mathbb{N}.$$

By the same arguments as in Theorem 2.1, and choosing  $\gamma$  as in the assertion of the theorem, we obtain that there exists  $K$  (only depending upon  $K_F, \beta$ , and  $\Delta_{\mathcal{X}}$ ) such that

$$\sum_{t=0}^{\infty} \beta^t \delta_t(\xi, \xi') \leq K (\mathbb{E} [d(\xi, \xi')])^\gamma, \quad \xi, \xi' \in L^0(\mathcal{G}; \mathcal{X}),$$

and we conclude with (4.4).  $\square$

**Theorem 4.2** *Assume that  $\mathcal{G}$  is rich enough and  $(\mathbf{H}_{\text{lip}})$  holds true. Then, for any  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ ,  $V(\xi) = \tilde{V}(\mu)$ , where  $\mu = \mathcal{L}(\xi)$ . Consequently,  $V$  is law-invariant, identified with  $\tilde{V}$ , and satisfies the Bellman fixed point equation  $\tilde{V} = \mathcal{T}\tilde{V}$ . Moreover, for all  $\epsilon > 0$ , there exists an  $\epsilon$ -optimal randomized feedback control for  $V(\xi)$ .*

**Proof.** As  $\mathcal{X}$  is compact, there exists a finite subset  $\mathcal{X}_\eta \subset \mathcal{X}$  such that  $d(x, x_\eta) \leq \eta$  for all  $x \in \mathcal{X}$ , where  $x_\eta$  denotes the projection of  $x$  on  $\mathcal{X}_\eta$ . Fix  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ , and set  $\mu = \mathcal{L}(\xi)$ . Let us then consider a random variable  $\xi' \sim \mu$  defined on another probability universe along with an independent uniform law  $U'$ . We set  $\Gamma' := (\xi', U')$ , and  $\mathcal{G}' = \sigma(\Gamma')$ . By construction the randomization hypothesis **Rand**( $\xi', \mathcal{G}'$ ) holds true, and we then have  $V(\xi') = \tilde{V}(\mu)$  from Theorem 4.1. Consider now the quantized random variables  $\xi_\eta$  and  $\xi'_\eta$ , which have the same law, and satisfy respectively the randomization hypothesis **Rand**( $\xi_\eta, \mathcal{G}$ ) and **Rand**( $\xi'_\eta, \mathcal{G}'$ ) from Lemma 3.2. From Theorem 4.1, we deduce that  $V(\xi_\eta) = \tilde{V}(\mathcal{L}(\xi_\eta)) = V(\xi'_\eta)$ . By uniform continuity of  $V$ , it follows by sending  $\eta$  to zero, that  $V(\xi) = V(\xi')$ , and thus  $V(\xi) = \tilde{V}(\mu)$ , which proves the required result.

Finally, the existence of an  $\epsilon$ -optimal control for  $V(\xi)$  is obtained as follows. From the uniform continuity of  $V$  in Lemma 4.6, there exists  $\eta$  small enough so that  $|V(\xi) - V(\xi_\eta)| \leq \epsilon/2$ . We then build according to Theorem 4.1 an  $\epsilon/2$ -optimal control for  $V(\xi_\eta)$ , which yields an  $\epsilon$ -optimal (randomized feedback stationary) control for  $V(\xi)$ .  $\square$

**Remark 4.9** From Theorems 4.1 and 4.2, under the condition that  $\mathcal{G}$  is rich enough and **(H<sub>lip</sub>)** holds true, the value function  $V$  of the CMKV-MDP is law-invariant, and the supremum in the Bellman fixed point equation for  $V \equiv \tilde{V}$  with the operator  $\mathcal{T}$  can be restricted to lifted randomized feedback policies, i.e.,

$$V = \mathcal{T}V = \sup_{\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)} \mathcal{T}^{\mathfrak{a}}V$$

where we set  $\mathcal{T}^{\mathfrak{a}} := \mathcal{T}^{\pi^{\mathfrak{a}}}$  equal to

$$[\mathcal{T}^{\mathfrak{a}}W](\mu) = \mathbb{E} \left[ f(Y^{\mathfrak{a}}(\mu, \xi, U)) + \beta W(\mathbb{P}_{F(Y^{\mathfrak{a}}(\mu, \xi, U), \varepsilon_1, \varepsilon_1^0)}^0) \right],$$

with  $Y^{\mathfrak{a}}(\mu, x, u) := (x, \mathfrak{a}(\mu, x, u), \pi^{\mathfrak{a}}(\mu))$ , and  $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ . Notice that this Bellman fixed point equation is not the same as the Bellman fixed point equation obtained by optimizing over feedback controls only (not randomized nor open-loop). Let us call  $V_f$  the associated optimal value. Then it is known that

$$V_f = \mathcal{T}_f V_f = \sup_{\mathfrak{a}_f \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X}; A)} \mathcal{T}^{\mathfrak{a}_f} V.$$

In other words, in the feedback case, the sup in the Bellman fixed point equation is only taken over (non-randomized) feedback policies.  $\square$

## 4.5 Feedback vs open-loop vs randomized controls

In this paper, we have mentioned different types of controls: open-loop controls, feedback controls, and randomized feedback controls. To fix ideas, let us consider a CMKV-MDP where **Rand**( $\xi, \mathcal{G}$ ) does not initially hold true, and let us address three problems:

- **Feedback problem:** Optimizing over stationary feedback controls. We note  $V_f$  the corresponding optimal value.
- **Open-loop problem:** Optimizing over open-loop controls. We note  $V_{ol}$  the corresponding optimal value.

- **Randomized feedback problem:** Enlarging  $\mathcal{G}$  with an independent uniform random variable  $U$ , i.e.  $\tilde{\mathcal{G}} := \sigma(\mathcal{G}, U)$ , so that **Rand**( $\xi, \tilde{\mathcal{G}}$ ) holds true, and then optimizing over stationary randomized feedback controls. We note  $V_r$  the corresponding optimal value.

We shall compare these problems by relying on the results of this paper as well as the examples presented at the end of this section. We will focus on the following questions:

- (1) **When is Bellman fixed point equation satisfied?** Theorem 4.1 shows that  $V_r$  satisfies a Bellman equation, and we have already mentioned in Remark 4.9 that  $V_f$  does satisfy a Bellman fixed point equation as well (although not the same). Regarding the open-loop problem, we know that if  $\mathcal{G}$  is rich enough,  $V_{ol}$  satisfies the same Bellman equation as  $V_r$ . However, when  $\mathcal{G}$  is not rich enough, Example 4.2 below provides a counter-example, where the Bellman fixed point equation is not satisfied.
- (2) **When do the optimal values coincide?** Theorem 4.1 shows that  $V_r$  is the same as the optimal value when the optimization is performed over open-loop controls (with initial information  $\tilde{\mathcal{G}}$ ), and thus  $V_r(\xi) \geq V_{ol}(\xi)$  (because  $\mathcal{G} \subset \tilde{\mathcal{G}}$ ). Also, we clearly have  $V_{ol}(\xi) \geq V_f(\xi)$ . The problem is now to figure out if the inequalities can be strict. If  $\mathcal{G}$  is not rich enough, Examples 4.1, 4.2, and 4.3 below illustrate that one can have  $V_r(\xi) > V_{ol}(\xi) > V_f(\xi)$ . When  $\mathcal{G}$  is rich enough, Theorem 4.2 shows that  $V_r(\xi) = V_{ol}(\xi)$ . However, regarding the comparison with  $V_f$ , it is remarkable that Examples 4.4 and 4.5 below illustrate that the inequality  $V_r = V_{ol} > V_f$  can be strict even when  $\mathcal{G}$  is rich enough, and actually even when  $\xi$  is atomless.
- (3) **When does an optimal control exist?** Given that in the comparison between the randomized feedback problem and the open-loop problem, we have seen that the optimal values are equal as soon as  $\mathcal{G}$  is rich enough, it is natural to wonder if there is still a qualitative difference between both cases. It is notable that Examples 4.6 and 4.7 below illustrate that an optimal control can exist in the randomized feedback case but not in the open-loop case. Furthermore, the optimal control from the randomized feedback case can have a much simpler and more regular form than any  $\epsilon$ -optimal control from the open-loop case.

We conclude this section by a set of simple examples illustrating the subtle differences between each type of problem.

**Example 4.1 (Feedback problem,  $\mathcal{G}$  not rich enough)** *Let us take an example similar to Example 3.1 in [12]. Consider  $\mathcal{X} = \{-1, 1\} = A$ ,  $\varepsilon_1 \sim \mathcal{B}(1/2)$ ,  $F(x, a, \nu, e, e^0) = ax$ ,  $f(x, a, \nu) = -\mathcal{W}(pr_1 \star \nu, \mathcal{B}(1/2))$ . In other words, the reward is maximal and equal to 0 when the law of the state is a Bernoulli(1/2) on  $\mathcal{X}$ , and minimal equal to  $-1/2$  when the law of the state is a Dirac ( $\delta_{-1}$  or  $\delta_1$ ). Assume that  $\Gamma = 1$  a.s., so that that  $\mathcal{G}$  is the trivial  $\sigma$ -algebra. In this case,  $\xi =: x$  is necessarily deterministic. We perform the optimization over feedback controls. It is clear that the law of  $X_t$  will always be a Dirac, and thus the gain will be  $V_f(\xi) = \sum_{t=0}^{\infty} \beta^t (-\frac{1}{2}) = -\frac{1}{2(1-\beta)}$  which is the worst possible gain.*

**Example 4.2 (Open-loop problem,  $\mathcal{G}$  not rich enough)** *Let us consider the same problem as in Example 4.1, but optimize over open-loop controls instead.  $\mathcal{G}$  being the trivial  $\sigma$ -algebra,  $\xi =: x$  and  $\alpha_0$  are then necessarily deterministic, and thus  $X_1^{\xi, \alpha} = \alpha_0 \xi$  has to be deterministic as well, which yields rewards at  $t = 0, 1$  both equal to  $-\frac{1}{2}$ . By choosing a control  $\alpha_0 = 1$ ,  $\alpha_1 = \varepsilon_1$  and  $\alpha_t = 1$  afterwards, we*



have  $\mathbb{P}_{X_t}^0 = \delta_x$  for  $t = 0, 1$ , and  $\mathbb{P}_{X_t}^0 = \mathcal{B}(1/2)$  afterwards. This control is clearly optimal, and the associated gain is  $V_{ol}(\xi) = -\frac{1+\beta}{2}$ . Notice that in this example, where  $\mathcal{G}$  is not rich enough,  $V$  does not satisfy a Bellman equation. Indeed, if it did, we would have  $V_{ol}(x) = \sup_{a \in A} (-1/2 + \beta V_{ol}(ax))$ , which is equivalent to  $-\frac{1+\beta}{2} = -\frac{1}{2} - \beta \frac{1+\beta}{2}$ , and this is clearly false.

**Example 4.3 (Randomized feedback problem,  $\mathcal{G}$  not rich enough)** Let us consider the same problem as in Example 4.1, but let us enlarge  $\mathcal{G}$  with an independent uniform random variable  $U$ , i.e.  $\tilde{\mathcal{G}} := \sigma(\mathcal{G}, U)$ , so that **Rand** $(\xi, \tilde{\mathcal{G}})$  holds true, and then optimizing over stationary randomized feedback controls. The enlarged initial information allows to set  $\alpha_0 = \text{sgn}(U - \frac{1}{2})$  and  $\alpha_t = 1$  for  $t \in \mathbb{N}_*$ . It is clear that the strategy is optimal and leads to a gain  $V_r(\xi) = -\frac{1}{2}$ .

**Example 4.4 (Feedback problem,  $\mathcal{G}$  rich enough,  $\xi$  atomless)** Consider  $\mathcal{X} = \{-1, 1\} \cup [2, 3]$ ,  $A = \{-1, 1\}$ ,  $\varepsilon_1 \sim \mathcal{B}(1/2)$ ,  $F(x, a, \nu, e, e^0) = ax \mathbf{1}_{x \in \{-1, 1\}} + \mathbf{1}_{x \in [2, 3]}$ ,  $f(x, a, \nu) = -\mathcal{W}(pr_1 \star \nu, \mathcal{B}(1/2))$ . Assume that  $\Gamma \sim \mathcal{U}([0, 1]^2)$  and  $\xi = 2 + \Gamma$  a.s.. In particular,  $\mathcal{G}$  is rich enough and  $\xi$  is atomless. This is the same example as Example 4.3 except for the choice of  $\Gamma$  and  $\xi$ , and for the fact that we have extended  $F$  on  $\mathcal{X} = \{-1, 1\} \cup [2, 3]$ . Assume that we optimize over feedback controls. Given the definition of  $F$ , it is clear that at time  $t = 0$ , the control has no impact, and that  $X_1 = 1$  a.s.. From  $t = 1$ , we thus fall back into the problem from Example 4.1 and thus the optimal value is  $V_f(\xi) = -\mathcal{W}(\mathcal{U}([2, 3]), \mathcal{B}(1/2)) - \beta \frac{1+\beta}{2}$ .

**Example 4.5 (Open-loop problem,  $\mathcal{G}$  rich enough,  $\xi$  atomless)** Let us consider the same problem as in Example 4.4, but let us optimize over open-loop controls instead. Again, at time  $t = 0$ , the control has no impact, and  $X_1 = 1$  a.s.. From  $t = 1$ , we fall back into the problem from Example 4.2 except that  $\mathcal{G}$  is now rich enough. By Theorem 4.2, the optimal value from time  $t = 1$  is thus the same as for the randomized feedback problem from Example 4.3, and thus the optimal value is  $V_{ol}(\xi) = -\mathcal{W}(\mathcal{U}([2, 3]), \mathcal{B}(1/2)) - \beta \frac{1}{2}$ .

**Example 4.6 (Randomized feedback problem, existence of an optimal control)** Assume that  $\mathcal{X} = [-1, 1]$ ,  $A = \{-1, 1\}$ ,  $\varepsilon_1 \sim \mathcal{U}([-1, 1])$ ,  $F(x, a, \nu, e, e^0) = ax$ ,  $f(x, a, \nu) = -\mathcal{W}(pr_1 \star \nu, \mathcal{U}([-1, 1]))$ ,  $\Gamma = (U, W) \sim \mathcal{U}([0, 1]^2)$ , and  $\xi = W$ . Then **Rand** $(\xi, \mathcal{G})$  holds true. Similarly to Example 4.3, we set  $\alpha_0 = \text{sgn}(U - \frac{1}{2})$  and  $\alpha_t = 1$  for  $t \in \mathbb{N}_*$ . It is straightforward to show that the strategy is optimal and also leads to a gain  $V_r(\xi) = -\frac{1}{2}$ .

**Example 4.7 (Open-loop problem, no optimal control)** Let us take the same problem as in Example 4.6, but with  $\Gamma = W$ . Then **Rand** $(\xi, \mathcal{G})$  does not hold true anymore. Indeed, any  $\mathcal{G}$ -measurable variable independent from  $\xi$  is necessarily deterministic and therefore cannot be uniform on  $[0, 1]$ . By Theorem 4.2, as  $\mathcal{G}$  is rich enough, we know that we will also have  $V_{ol}(\xi) = -\frac{1}{2}$ . However, without the additional uniform variable  $U$ , the existence of an optimal control does not hold true anymore. Indeed, from Example 4.6, an optimal strategy would imply  $X_1 \sim \mathcal{U}([-1, 1])$ . On the other hand, since  $\alpha_0$  is  $\mathcal{G}$ -measurable, it has to take the (non-randomized) form  $\alpha_0 = h(\xi)$  for  $h \in L^0(\mathcal{X}, A)$ . The latter implies that the support of the distribution of  $X_1 = h(\xi)\xi$  is of Lebesgue measure 1, which is in contradiction with the  $\mathcal{U}([-1, 1])$ -distribution of  $X_1$ . Moreover, from the proof of Theorem 4.2, it is clear that building an approximate optimal control would require to introduce a quantized version  $\xi_\eta$  of  $\xi$  for  $\eta$  small, and then to use Lemma 3.2 to build a uniform variable  $U_\eta \perp \xi_\eta$ . Therefore, the approximate optimal control would be much more complex than the simple optimal control from Example 4.6.

## 4.6 Computing value function and $\epsilon$ -optimal strategies in CMKV-MDP

Having established the correspondence of our CMKV-MDP with lifted MDP on  $\mathcal{P}(\mathcal{X})$ , and the associated Bellman fixed point equation, we can (up to a simple discretization of the state space in the Bellman fixed point equation) design two methods for computing the value function and optimal strategies:

**(a) Value iteration.** We approximate the value function  $V = \tilde{V} = V^*$  by iteration from the Bellman operator:  $V_{n+1} = \mathcal{T}V_n$ , and at iteration  $N$ , we compute an approximate optimal randomized feedback policy  $\mathbf{a}_N$  by (recall Remark 4.9)

$$\mathbf{a}_N \in \arg \max_{\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)} \mathcal{T}^{\mathbf{a}} V_N.$$

From  $\mathbf{a}_N$ , we then construct an approximate randomized feedback stationary control  $\alpha^{\mathbf{a}_N}$  according to the procedure described in Remark 4.4.

**(b) Policy iteration.** Starting from some initial randomized feedback policy  $\mathbf{a}_0 \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$ , we iterate according to:

- Policy evaluation: we compute the expected gain  $\tilde{V}^{\pi^{\mathbf{a}_0}}$  of the lifted MDP
- Greedy strategy: we compute

$$\mathbf{a}_{k+1} \in \arg \max_{\mathbf{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)} \mathcal{T}^{\mathbf{a}} \tilde{V}^{\pi^{\mathbf{a}_k}}.$$

We stop at iteration  $K$  to obtain  $\mathbf{a}_K$ , and then construct an approximate randomized feedback control  $\alpha^{\mathbf{a}_K}$  according to the procedure described in Remark 4.4.

**Practical computation.** Since a randomized feedback control  $\alpha$  is a measurable function  $\mathbf{a}$  of  $(\mathbb{P}_{X_t^{\xi, \alpha}}^0, X_t^{\xi, \alpha}, U_t)$ , we would need to compute and store the (conditional) law of the state process, which is infeasible in practice when  $\mathcal{X}$  is a continuous space. In this case, to circumvent this issue, a natural idea is to discretize the compact space  $\mathcal{X}$  by considering a finite subset  $\mathcal{X}_\eta = \{x^1, \dots, x^{N_\eta}\} \subset \mathcal{X}$  associated with a partition  $B_\eta^i$ ,  $i = 1, \dots, N_\eta$ , of  $\mathcal{X}$ , satisfying:  $B_\eta^i \subset \{x \in \mathcal{X} : d(x, x^i) \leq \eta\}$ ,  $i = 1, \dots, N_\eta$ , with  $\eta > 0$ . For any  $x \in \mathcal{X}$ , we denote by  $[x]_\eta$  (or simply  $x_\eta$ ) its projection on  $\mathcal{X}_\eta$ , defined by:  $x_\eta = x^i$ , for  $x \in B_\eta^i$ ,  $i = 1, \dots, N_\eta$ .

**Definition 4.2 (Discretized CMKV-MDP)** Fix  $\eta > 0$ . Given  $\xi \in L^0(\mathcal{G}; \mathcal{X}_\eta)$ , and a control  $\alpha \in \mathcal{A}$ , we denote by  $X^{\eta, \xi, \alpha}$  the McKean-Vlasov MDP on  $\mathcal{X}_\eta$  given by

$$X_{t+1}^{\eta, \xi, \alpha} = [F(X_t^{\eta, \xi, \alpha}, \alpha_t, \mathbb{P}_{(X_t^{\eta, \xi, \alpha}, \alpha_t)}^0, \varepsilon_{t+1}, \varepsilon_{t+1}^0)]_\eta, \quad t \in \mathbb{N}, \quad X_0^{\eta, \xi, \alpha} = \xi,$$

i.e., obtained by projecting the state on  $\mathcal{X}_\eta$  after each application of the transition function  $F$ . The associated expected gain  $V_\eta^\alpha$  is defined by

$$V_\eta^\alpha(\xi) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t f(X_t^{\eta, \xi, \alpha}, \alpha_t, \mathbb{P}_{(X_t^{\eta, \xi, \alpha}, \alpha_t)}^0) \right].$$

Notice that the (conditional) law of the discretized CMKV-MDP on  $\mathcal{X}_\eta$  is now valued in a finite-dimensional space (the simplex of  $[0, 1]^{N_\eta}$ ), which makes the computation of the associated randomized

feedback control accessible, although computationally challenging due to the high-dimensionality (and beyond the scope of this paper). The next result states that an  $\epsilon$ -optimal randomized feedback control in the initial CMKV-MDP can be approximated by a randomized feedback control in the discretized CMKV-MDP.

**Proposition 4.4** *Assume that  $\mathcal{G}$  is rich enough and  $(\mathbf{H}_{\text{lip}})$  holds true. Fix  $\xi \in L^0(\mathcal{G}; \mathcal{X})$ . Given  $\eta > 0$ , let us define  $\xi_\eta$  the projection of  $\xi$  on  $\mathcal{X}_\eta$ . As  $\mathbf{Rand}(\xi_\eta, \mathcal{G})$  holds true, let us consider an i.i.d. sequence  $(U_{\eta,t})_{t \in \mathbb{N}}$  of  $\mathcal{G}$ -measurable uniform variables independent of  $\xi_\eta$ . For  $\epsilon > 0$ , let  $\mathbf{a}_\epsilon$  be a randomized feedback policy that is  $\epsilon$ -optimal for the Bellman fixed point equation satisfied by  $V$ . Finally, let  $\alpha^{\eta, \epsilon}$  be the randomized feedback control in the discretized CMKV-MDP recursively defined by  $\alpha_t^{\eta, \epsilon} = \mathbf{a}_\epsilon(\mathbb{P}_{X_t^{\eta, \epsilon}}^0, X_t^{\eta, \epsilon}, U_{\eta,t})$ ,  $t \in \mathbb{N}$ , where we set  $X_t^{\eta, \epsilon} := X_t^{\eta, \xi_\eta, \alpha^{\eta, \epsilon}}$ . Then the control  $\alpha^{\eta, \epsilon}$  is  $\mathcal{O}(\eta^\gamma + \epsilon)$ -optimal for the CMKV-MDP  $X$  with initial state  $\xi$ , where  $\gamma = \min(1, \frac{|\ln \beta|}{(\ln 2K)_+})$ .*

**Proof.** *Step 1.* Let us show that

$$\sup_{\alpha \in \mathcal{A}} \sum_{t=0}^{\infty} \beta^t \mathbb{E}[d(X_t^{\xi, \alpha}, X_t^{\eta, \xi_\eta, \alpha})] \leq C\eta^\gamma, \quad (4.23)$$

for some constant  $C$  that depends only on  $K$ ,  $\beta$  and  $\gamma$ . Indeed, notice by definition of the projection on  $\mathcal{X}_\eta$ , and by a simple conditioning argument that for all  $\alpha \in \mathcal{A}$ , and  $t \in \mathbb{N}$ ,

$$\mathbb{E}[d(X_{t+1}^{\xi, \alpha}, X_{t+1}^{\eta, \xi_\eta, \alpha})] \leq \eta + \mathbb{E}[\Delta(X_t^{\xi, \alpha}, X_t^{\eta, \xi_\eta, \alpha}, \alpha_t, \mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \mathbb{P}_{(X_t^{\eta, \xi_\eta, \alpha}, \alpha_t)}^0, \varepsilon_{t+1}^0)],$$

where

$$\Delta(x, x', a, \nu, \nu', e^0) = \mathbb{E}[d(F(x, a, \nu, \varepsilon_{t+1}, e^0), F(x', a, \nu', \varepsilon_{t+1}, e^0))].$$

Under  $(\mathbf{H}_{\text{lip}})$ , we then get

$$\begin{aligned} \mathbb{E}[d(X_{t+1}^{\xi, \alpha}, X_{t+1}^{\eta, \xi_\eta, \alpha})] &\leq \eta + K\mathbb{E}\left[d(X_t^{\xi, \alpha}, X_t^{\eta, \xi_\eta, \alpha}) + \mathcal{W}(\mathbb{P}_{X_t^{\xi, \alpha}}^0, \mathbb{P}_{X_t^{\eta, \xi_\eta, \alpha}}^0)\right] \\ &\leq \eta + 2K\mathbb{E}[d(X_t^{\xi, \alpha}, X_t^{\eta, \xi_\eta, \alpha})], \end{aligned}$$

by the same argument as in (4.4). Hence, the sequence  $(\mathbb{E}[d(X_t^{\xi, \alpha}, X_t^{\eta, \xi_\eta, \alpha})])_{t \in \mathbb{N}}$  satisfies the same type of induction inequality as in (2) in Theorem 2.1 with  $\eta$  instead of  $M_N$ , and thus the same derivation leads to the required result (4.6). From the Lipschitz condition on  $f$ , we deduce by the same arguments as in (4.4) in Lemma 4.6 that

$$\sup_{\alpha \in \mathcal{A}} |V^\alpha(\xi_\eta) - V_\eta^\alpha(\xi_\eta)| = \mathcal{O}(\eta^\gamma). \quad (4.24)$$

*Step 2.* Denote by  $\mu = \mathcal{L}(\xi)$ , and  $\mu_\eta = \mathcal{L}(\xi_\eta)$ , and observe that  $\mathcal{W}(\mu, \mu_\eta) \leq \mathbb{E}[d(\xi, \xi_\eta)] \leq \eta$ . We write

$$\begin{aligned} V^{\alpha^{\eta, \epsilon}}(\xi) - V(\xi) &= [V^{\alpha^{\eta, \epsilon}}(\xi) - V^{\alpha^{\eta, \epsilon}}(\xi_\eta)] + [V^{\alpha^{\eta, \epsilon}}(\xi_\eta) - V_\eta^{\alpha^{\eta, \epsilon}}(\xi_\eta)] \\ &\quad + [V_\eta^{\alpha^{\eta, \epsilon}}(\xi_\eta) - V(\xi_\eta)] + [V(\xi_\eta) - V(\xi)] =: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

The first and last terms  $I_1$  and  $I_4$  are smaller than  $\mathcal{O}(\eta^\gamma)$  by the  $\gamma$ -Hölder property of  $V^\alpha$  and  $V$  in Lemma 4.6. By (4.6), the second term  $I_2$  is of order  $\mathcal{O}(\eta^\gamma)$  as well for  $\eta$  small enough. Regarding the third term  $I_3$ , notice that by definition,  $V_\eta^{\alpha^{\eta, \epsilon}}(\xi_\eta)$  corresponds to the gain associated to the randomized

feedback policy  $\mathbf{a}_\epsilon$  for the discretized CMKV-MDP. Denote by  $\boldsymbol{\pi}_\epsilon$  the lifted randomized feedback policy associated to  $\mathbf{a}_\epsilon$ , and recall by Remark 4.4 the identification with the lifted MDP:  $V_\eta^{\alpha^{\eta,\epsilon}}(\xi') = \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu')$ ,  $\mu' = \mathcal{L}(\xi')$ , where  $\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}$  is the expected gain of the lifted MDP associated to the discretized CMKV-MDP, hence fixed point of the operator

$$[\mathcal{T}_\eta^{\mathbf{a}_\epsilon} W](\mu') = \mathbb{E} \left[ f(Y^{\mathbf{a}_\epsilon}(\mu', \xi', U)) + \beta W \left( \mathbb{P}_{[F(Y^{\mathbf{a}_\epsilon}(\mu', \xi', U), \varepsilon_1, \varepsilon_1^0)]_\eta}^0 \right) \right],$$

$Y^\mathbf{a}(\mu, x, u) = (x, \mathbf{a}(\mu, x, u), \boldsymbol{\pi}^\mathbf{a}(\mu))$  and  $(\xi', U) \sim \mu' \otimes \mathcal{U}([0, 1])$ . Recalling that  $V(\xi') = \tilde{V}(\mu')$ ,  $\mu' = \mathcal{L}(\xi')$ , with  $\tilde{V}$  fixed point to the Bellman operator  $\mathcal{T}$ , it follows that

$$\begin{aligned} I_3 = \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu_\eta) - \tilde{V}(\mu_\eta) &= \left( [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}](\mu_\eta) - [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) \right) + \left( [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) - [\mathcal{T}^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) \right) \\ &+ \left( [\mathcal{T}^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) - \tilde{V}(\mu_\eta) \right) =: I_3^1 + I_3^2 + I_3^3. \end{aligned}$$

By definition of  $\mathbf{a}_\epsilon$ , we have  $|I_3^3| \leq \epsilon$ . For  $I_3^2$  notice that the only difference between the operators  $\mathcal{T}_\eta^{\mathbf{a}_\epsilon}$  and  $\mathcal{T}^{\mathbf{a}_\epsilon}$  is that  $F$  is projected on  $\mathcal{X}_\eta$ . Thus,

$$\left| [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) - [\mathcal{T}^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) \right| \leq \beta \mathbb{E} \left[ \left| \tilde{V} \left( \mathbb{P}_{[F(Y_\eta, \varepsilon_1, \varepsilon_1^0)]_\eta}^0 \right) - \tilde{V} \left( \mathbb{P}_{F(Y_\eta, \varepsilon_1, \varepsilon_1^0)}^0 \right) \right| \right],$$

where  $Y_\eta = (\xi_\eta, \mathbf{a}_\epsilon(\mu, \xi_\eta, U), \boldsymbol{\pi}_\epsilon(\mu_\eta))$ . It is clear by definition of the Wasserstein distance and the projection on  $\mathcal{X}_\eta$  that

$$\mathcal{W} \left( \mathbb{P}_{[F(Y_\eta, \varepsilon_1, \varepsilon_1^0)]_\eta}^0, \mathbb{P}_{F(Y_\eta, \varepsilon_1, \varepsilon_1^0)}^0 \right) \leq \mathbb{E}^0 [d(F(Y_\eta, \varepsilon_1, \varepsilon_1^0), [F(Y_\eta, \varepsilon_1, \varepsilon_1^0)]_\eta)] \leq \eta.$$

From the  $\gamma$ -Hölder property of  $\tilde{V}$  in Proposition 4.2, we deduce that  $I_3^2 = \mathcal{O}(\eta^\gamma)$ . Finally, for  $I_3^1$ , since  $\mathcal{T}_\eta^{\mathbf{a}_\epsilon}$  is a  $\beta$ -contracting operator on  $(L^\infty(\mathcal{M}_\eta), \|\cdot\|_{\eta, \infty})$ , we have

$$\left| [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}](\mu_\eta) - [\mathcal{T}_\eta^{\mathbf{a}_\epsilon} \tilde{V}](\mu_\eta) \right| \leq \beta \|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta, \infty},$$

and thus  $|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu_\eta) - \tilde{V}(\mu_\eta)| = |I_3| \leq |I_3^1| + |I_3^2| + |I_3^3| \leq \beta \|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta, \infty} + \mathcal{O}(\eta^\gamma + \epsilon)$ . Taking the sup over  $\mu_\eta \in \mathcal{M}_\eta$  on the left, we obtain that  $\|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta, \infty} \leq \frac{1}{1-\beta} \mathcal{O}(\eta^\gamma + \epsilon) = \mathcal{O}(\eta^\gamma + \epsilon)$ , and we conclude that  $|I_3| \leq \|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta, \infty} \leq \mathcal{O}(\eta^\gamma + \epsilon)$ , which ends the proof.  $\square$

**Remark 4.10** Back to the  $N$ -agent MDP problem with open-loop controls, recall from Section 2, that it suffices to find an  $\epsilon$ -optimal open-loop policy  $\pi^\epsilon \in \Pi_{OL}$  for the CMKV-MDP, as it will automatically be  $\mathcal{O}(\epsilon)$ -optimal for the  $N$ -agent MDP with  $N$  large enough. For instance, the construction of an  $\epsilon$ -optimal control  $\alpha^\epsilon$  given by Proposition 4.4 can be associated to an  $\epsilon$ -optimal open-loop policy  $\pi^\epsilon$  such that  $\alpha^\epsilon = \alpha^{\pi^\epsilon}$  (where  $\pi_t^\epsilon$  is a measurable function of  $(\Gamma, (\varepsilon_s)_{s \leq s \leq t}, (\varepsilon_s^0)_{s \leq s \leq t})$ ). The processes  $\mathcal{O}(\epsilon)$ -optimal for the  $i$ -th agent  $\alpha_t^{\varepsilon^i}$  is then the result of the same construction but with  $(\Gamma^i, \varepsilon^i, \varepsilon^0)$  instead of  $(\Gamma, \varepsilon, \varepsilon^0)$ , i.e. replacing  $\xi$  by  $\xi^i$ ,  $U_{\eta, t}$  by  $U_{\eta, t}^i$ , and  $(\Gamma, \varepsilon, \varepsilon^0)$  by  $(\Gamma^i, \varepsilon^i, \varepsilon^0)$  in Proposition 4.4. Notice that this construction never requires the access to the individual's states  $X^{i, N}$ .

**Remark 4.11 ( $Q$  function)** In view of the Bellman fixed point equation satisfied by the value function  $V$  of the CMKV-MDP in terms of randomized feedback policies, let us introduce the corresponding state-action value function  $Q$  defined on  $\mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$  by

$$Q(\mu, \hat{a}) = [\hat{\mathcal{T}}^{\hat{a}} V](\mu) = \hat{f}(\mu, \hat{a}) + \beta \mathbb{E} [V(\hat{F}(\mu, \hat{a}, \varepsilon_1^0))],$$

From Proposition 4.1, and since  $V = \mathcal{T}V$ , we recover the standard connection between the value function and the state-action value function, namely  $V(\mu) = \sup_{\hat{a} \in \hat{A}(\mathcal{X})} Q(\mu, \hat{a})$ , from which we obtain the Bellman equation for the  $Q$  function:

$$Q(\mu, \hat{a}) = \hat{f}(\mu, \hat{a}) + \beta \mathbb{E} \left[ \sup_{\hat{a}' \in \hat{A}(\mathcal{X})} Q(\mu_1^{\hat{a}}, \hat{a}') \right], \quad (4.25)$$

where we set  $\mu_1^{\hat{a}} = \hat{F}(\mu, \hat{a}, \varepsilon_1^0)$ . Notice that this  $Q$ -Bellman equation extends the equation in [12] (see their Theorem 3.1) derived in the no common noise case and when there is no mean-field dependence with respect to the law of the control. The Bellman equation (4.11) is the starting point in a model-free framework when the state transition function is unknown (in other words in the context of reinforcement learning) for the design of  $Q$ -learning algorithms in order to estimate the  $Q$ -value function by  $Q_n$ , and then to compute a relaxed control by

$$\hat{a}_n^\mu \in \arg \max_{\hat{a} \in \hat{A}(\mathcal{X})} Q_n(\mu, \hat{a}), \quad \mu \in \mathcal{P}(\mathcal{X}).$$

From Lemma 2.22 [14], one can associate to such probability kernel  $\hat{a}_n^\mu$ , a function  $\mathbf{a}_n : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1] \rightarrow A$ , such that  $\mathcal{L}(\mathbf{a}_n(\mu, x, U)) = \hat{a}_n^\mu(x)$ ,  $\mu \in \mathcal{P}(\mathcal{X})$ ,  $x \in \mathcal{X}$ , where  $U$  is an uniform random variable. In practice, one has to discretize the state space  $\mathcal{X}$  as in Definition 4.2, and then to quantize the space  $\mathcal{P}(\mathcal{X})$  as in Lemma 4.5 in order to reduce the learning problem to a finite-dimensional problem for the computation of an approximate optimal randomized feedback policy  $\mathbf{a}_n$  for the CMKV-MDP.  $\square$

**Remark 4.12 (About the practical implementation of an RL algorithm)** The individualized and open-loop form of controls has been motivated in the introduction from the point of view of applications. Let us now discuss some implications regarding the practical implementation of algorithms. As described in the introduction, the algorithms for computing the actions will be running in parallel on each individual's device (phone or computer), and in order to reduce the cost of data transfer, they will only access the data available on their respective devices. For instance, on device  $i$ , only the individual's own data  $\varepsilon^i$  and the public data  $\varepsilon^0$  will be accessible. These data are used to compute the McKean-Vlasov process  $X^i$  associated to the individual. A randomized feedback policy is then applied to  $X^i$  to compute the action  $\alpha^i$ .

An important consequence is that each algorithm must be able to compute the McKean-Vlasov process. Therefore: (1) Either the dynamic  $F$  and reward  $f$  are explicitly modeled. (2) Either the population's behavior is simulated without explicitly describing  $F$  and  $f$ . The first case naturally leads to MDP algorithms (value/policy iteration), while the second case is better suited for RL algorithms. We refer to the updated book [21] for a general overview of RL algorithms, and to the more recent papers [6] and [13] in the framework of mean-field control/games.

## 5 Conclusion

We have developed a theory for mean-field Markov decision processes with common noise and open-loop controls, called CMKV-MDP, for general state space and action space. Such problem is motivated and shown to be the asymptotic problem of a large population of cooperative agents under mean-field interaction controlled by a social planner/influencer, and we provide a rate of convergence of the  $N$ -agent model to the CMKV-MDP. We prove the correspondence of CMKV-MDP with a general lifted MDP on the space of probability measures, and emphasize the role of relaxed control, which is crucial

to characterize the solution via the Bellman fixed point equation. Approximate randomized feedback controls are obtained from the Bellman equation in a model-based framework, and future work under investigation will develop algorithms in a model-free framework, in other words in the context of reinforcement learning with many interacting and cooperative agents.

## A Some useful results on conditional law

**Lemma A.1** *Let  $(S, \mathcal{S})$ ,  $(T, \mathcal{T})$ , and  $(U, \mathcal{U})$  be three measurable spaces, and  $F \in L^0((S, \mathcal{S}) \times (T, \mathcal{T}); (U, \mathcal{U}))$  be a measurable function, then the function  $\hat{F} : (\mathcal{P}(S), \mathcal{C}(S)) \times (T, \mathcal{T}) \rightarrow (\mathcal{P}(U), \mathcal{C}(U))$  given by  $\hat{F}(\mu, x) := F(\cdot, x) \star \mu$  is measurable.*

**Proof.** This follows from the measurability of the maps:

- $x \in (S, \mathcal{S}) \mapsto \delta_x \in (\mathcal{P}(S), \mathcal{C}(S))$ ,
- $(\mu, \nu) \in (\mathcal{P}(S), \mathcal{C}(S)) \times (\mathcal{P}(T), \mathcal{C}(T)) \mapsto \mu \otimes \nu \in (\mathcal{P}(S \times T), \mathcal{C}(S \times T))$ ,
- $\mu \in (\mathcal{P}(S), \mathcal{C}(S)) \mapsto F \star \mu \in (\mathcal{P}(T), \mathcal{C}(T))$ ,

and the measurability of the composition  $(\mu, x) \mapsto (\mu, \delta_x) \mapsto \mu \otimes \delta_x \mapsto F \star (\mu \otimes \delta_x) = \hat{F}(\mu, x)$ .  $\square$

**Lemma A.2 (Conditional law)** *Let  $(S, \mathcal{S})$  and  $(T, \mathcal{T})$  be two measurable spaces.*

1. *If  $(S, \mathcal{S})$  is a Borel space, there exists a conditional law of  $Y$  knowing  $X$ .*
2. *If  $Y = \varphi(X, Z)$  where  $Z \perp X$  is a random variable valued in a measurable space  $V$  and  $\varphi : S \times V \rightarrow T$  is a measurable function, then  $\mathcal{L}(\varphi(x, Z))|_{x=X}$  is a conditional law of  $Y$  knowing  $X$ . In the case  $S = S_1 \times S_2$ ,  $X = (X_1, X_2)$ , and  $Y = \varphi(X_1, Z)$ , then  $\mathbb{P}_Y^X = \mathcal{L}(\varphi(x_1, Z))|_{x_1=X_1}$ , and thus  $\mathbb{P}_Y^X$  is  $\sigma(X_1)$ -measurable in  $(\mathcal{P}(T), \mathcal{C}(T))$ .*
3. *For any probability kernel  $\nu$  from  $S$  to  $\mathcal{P}(T)$ , there exists a measurable function  $\phi : S \times [0, 1] \rightarrow T$  s.t.  $\nu(s) = \mathcal{L}(\phi(s, U))$ , for all  $s \in S$ , where  $U$  is a uniform random variable.*

**Proof.** The first assertion is stated in Theorem 6.3 in [14], and the second one follows from Fubini's theorem. The third assertion is a consequence of the two others.  $\square$

**Proposition A.1** *Given an open-loop control  $\alpha \in \mathcal{A}$ , and an initial condition  $\xi \in L^0(\mathcal{X}; \mathcal{G})$ , the solution  $X^{\xi, \alpha}$  to the conditional McKean-Vlasov equation is such that: for all  $t \in \mathbb{N}$ ,  $X_t^{\xi, \alpha}$  is  $\sigma(\xi, \Gamma, (\varepsilon)_{s \leq t}, (\varepsilon_s^0)_{s \leq t})$ -measurable, and  $\mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0$  is  $\mathcal{F}_t^0$ -measurable.*

**Proof.** We prove the result by induction on  $t$ . It is clear for  $t = 0$ . Assuming that it holds true for some  $t \in \mathbb{N}$ , we write

$$X_{t+1}^{\xi, \alpha} = F(X_t^{\xi, \alpha}, \alpha_t, \mathbb{P}_{(X_t^{\xi, \alpha}, \alpha_t)}^0, \varepsilon_{t+1}, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}.$$

By induction hypothesis, there is a measurable function  $f_{t+1} : \mathcal{X} \times G \times E^{t+1} \times (E^0)^{t+1} \rightarrow \mathcal{X}$  s.t.  $X_{t+1}^{\xi, \alpha} = f_{t+1}(\xi, \Gamma, (\varepsilon_s)_{s \leq t+1}, (\varepsilon_s^0)_{s \leq t+1})$ , and thus  $X_{t+1}^{\xi, \alpha}$  is  $\sigma(\xi, \Gamma, (\varepsilon)_{s \leq t+1}, (\varepsilon_s^0)_{s \leq t+1})$ -measurable and  $\mathbb{P}_{(X_{t+1}^{\xi, \alpha}, \alpha_{t+1})}^0$  is  $\sigma(\varepsilon_s^0, s \leq t+1)$ -measurable by Lemma A.2.  $\square$

## B Wasserstein convergence of the empirical law

**Proposition B.1 (Conditional Wasserstein convergence of the empirical measure)** *Let  $E, F$  be two measurable spaces and  $G$  a compact Polish space. Let  $X$  be an  $E$ -valued random variable independent from a family of i.i.d.  $F$ -valued variables  $(U_i)_{i \in \mathbb{N}}$ , and a measurable function  $f : E \times F \rightarrow G$ . Then*

$$\mathcal{W}\left(\frac{1}{N} \sum_{i=1}^N \delta_{f(X, U_i)}, \mathbb{P}_{f(X, U_1)}^X\right) \xrightarrow[N \rightarrow \infty]{a.s.} 0.$$

**Proof.** It suffices to observe that the probability of this event is one by conditioning w.r.t.  $X$  and use the analog non-conditional result, which follows from the fact that Wasserstein distance metrizes weak convergence (as  $G$  is compact), and the fact that empirical measure converges weakly.  $\square$

## C Proof of coupling results

**Lemma C.1** *Let  $U, V$  be two independent uniform variables, and  $F$  a distribution function on  $\mathbb{R}$ . We have*

$$\left(F^{-1}(U), F(F^{-1}(U)) - U\right) \stackrel{d}{=} \left(F^{-1}(U), V \Delta F(F^{-1}(U))\right),$$

where we denote  $\Delta F := F - F_-$ .

**Proof.** Notice that  $F(F^{-1}(U)) - U$  is the position (from top to bottom) of  $U$  in the set  $\{u \in [0, 1], F^{-1}(u) = F^{-1}(U)\}$  and is thus smaller than  $\Delta F(F^{-1}(U))$ . Now, given a measurable function  $f \in L^0(A \times [0, 1]; \mathbb{R})$ , we have

$$\begin{aligned} & \mathbb{E}\left[f(F^{-1}(U), F(F^{-1}(U)) - U)\right] \\ &= \mathbb{E}\left[f(F^{-1}(U), 0) \mathbf{1}_{\Delta F(F^{-1}(U))=0}\right] + \mathbb{E}\left[f(F^{-1}(U), F(F^{-1}(U)) - U) \mathbf{1}_{\Delta F(F^{-1}(U))>0}\right]. \end{aligned} \tag{C.1}$$

The second term can be decomposed as

$$\sum_{\Delta F(c)>0} \mathbb{E}\left[f(c, F(c) - U) \mathbf{1}_{F^{-1}(U)=c}\right] = \sum_{\Delta F(c)>0} \int_0^1 f(c, \Delta F(c)u) \Delta F(c) du.$$

where the equality comes from a change of variable. Summing over  $\Delta F(c) > 0$ , we obtain  $\mathbb{E}[f(F^{-1}(U), V \Delta F(F^{-1}(U)))]$  and combined with (C), we get

$$\mathbb{E}\left[f(F^{-1}(U), F(F^{-1}(U)) - U)\right] = \mathbb{E}\left[f(F^{-1}(U), V \Delta F(F^{-1}(U)))\right],$$

which proves the result.  $\square$

**Lemma C.2** *Let  $\mathcal{X}$  be a compact Polish space, then there exists an embedding  $\phi \in L^0(\mathcal{X}, \mathbb{R})$  such that*

1.  $\phi$  and  $\phi^{-1}$  are uniformly continuous,

2. for any probability measure  $\mu \in \mathcal{P}(\mathcal{X})$ , we have  $\text{Im}(F_{\phi \star \mu}^{-1}) \subset \text{Im}(\phi)$ . In particular,  $\phi^{-1} \circ F_{\phi \star \mu}^{-1}$  is well posed.

**Proof.** 1. Without loss of generality, we assume that  $\mathcal{X}$  is bounded by 1. Fix a countable dense family  $(x_n)_{n \in \mathbb{N}}$  in  $\mathcal{X}$ . We define the map  $\phi_1 : x \in \mathcal{X} \mapsto (d(x, x_n))_{n \in \mathbb{N}} \in [0, 1]^{\mathbb{N}}$ . Let us endow  $[0, 1]^{\mathbb{N}}$  with the metric  $d((u_n)_{n \in \mathbb{N}}, (v_n)_{n \in \mathbb{N}}) := \sum_{n \geq 0} \frac{1}{2^n} |u_n - v_n|$ .  $\phi_1$  is clearly injective and uniformly continuous (even Lipschitz). The compactness of  $\mathcal{X}$  implies that its inverse  $\phi_1^{-1}$  is uniformly continuous as well. Let us now consider  $\phi_2 : ([0, 1]^{\mathbb{N}}, d) \mapsto [0, 1]$  where  $\phi_2((u_n)_{n \in \mathbb{N}})$  essentially groups the decimals of the real numbers  $u_n$ ,  $n \in \mathbb{N}$ , in a single real number. More precisely, let  $\iota : \mathbb{N} \rightarrow \mathbb{N}^2$  be a surjection, then we define the  $k$ -th decimal of  $\phi_2((u_n)_{n \in \mathbb{N}})$  as the  $(\iota(k))_2$ -th decimal of  $u_{(\iota(k))_1}$  (with the convention that for a number with two possible decimal representations, we choose the one that ends with 000...).  $\phi_2$  is clearly injective, uniformly continuous, as well as its inverse  $\phi_2^{-1}$ . Thus,  $\phi := \phi_2 \circ \phi_1$  defines an embedding of  $\mathcal{X}$  into  $\mathbb{R}$ , such that  $\phi$  and  $\phi^{-1}$  are uniformly continuous.

2.  $F_{\phi \star \mu}^{-1}$  being caglad, and  $\text{Im}(\phi)$  being closed (by compactness of  $\mathcal{X}$ ), it is enough to prove that  $F_{\phi \star \mu}^{-1}(u) \in \text{Im} \phi$  for almost every  $u \in [0, 1]$  (in the Lebesgue sense). However, given a uniform variable  $U$ , we have  $F_{\phi \star \mu}^{-1}(U) \sim \phi \star \mu$ , and thus

$$\mathbb{P}(F_{\phi \star \mu}^{-1}(U) \in \text{Im}(\phi)) = \mathbb{P}_{Y \sim \mu}(\phi(Y) \in \text{Im}(\phi)) = 1.$$

□

### Proof of Lemma 4.1

(1) We first consider the case where  $\mathcal{X} \subset \mathbb{R}$ . Let us call  $F_\mu$  the distribution function of  $\mu \in \mathcal{P}(\mathcal{X})$ , and  $F_\mu^{-1}$  its generalized inverse. Let us define the function  $\zeta : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0, 1] \rightarrow \mathcal{X}$  by

$$\zeta(\mu, \mu', x, u) := F_{\mu'}^{-1}(F_\mu(x) - u \Delta F_\mu(x)),$$

which is measurable by noting that the measurability in  $\mu, \mu'$  comes from the continuity of

$$\begin{aligned} \mathcal{P}(\mathcal{X}) &\rightarrow L_{caglad}^1([0, 1], \mathcal{X}) \\ \mu &\mapsto F_\mu^{-1}. \end{aligned}$$

By construction, we then have for any  $\xi \sim \mu$ , and  $U, V$  two independent uniform variables, independent of  $\xi$

$$\begin{aligned} (\xi, \zeta(\mu, \mu', \xi, V)) &= (\xi, F_{\mu'}^{-1}(F_\mu(\xi) - V \Delta F_\mu(\xi))) \\ &\stackrel{d}{=} (F_\mu^{-1}(U), F_{\mu'}^{-1}(F_\mu(F_\mu^{-1}(U)) - V \Delta F_\mu(F_\mu^{-1}(U)))) \\ &= (F_\mu^{-1}(U), F_{\mu'}^{-1}(F_\mu(F_\mu^{-1}(U)) - V \Delta F_\mu(F_\mu^{-1}(U)))) \\ &\stackrel{d}{=} (F_\mu^{-1}(U), F_{\mu'}^{-1}(U)), \end{aligned}$$

where the last equality holds by Lemma C.1. It is well-known (see e.g. Theorem 3.1.2 in [20]) that  $(F_\mu^{-1}(U), F_{\mu'}^{-1}(U))$  is an optimal coupling for  $(\mu, \mu')$ , and so  $\mathcal{W}(\mu, \mu') = \mathbb{E}[d(\xi, \zeta(\mu, \mu', \xi, V))]$ .

(2) Let us now consider the case of a general compact Polish space  $\mathcal{X}$ . Denoting by  $\zeta_{\mathbb{R}}$  the “ $\zeta$ ” from the case “ $\mathcal{X} \subset \mathbb{R}$ ”, and considering an embedding  $\phi \in L^0(\mathcal{X}, \mathbb{R})$  as in Lemma C.2, let us define

$$\zeta(\mu, \mu', x, u) := \phi^{-1}(\zeta_{\mathbb{R}}(\phi \star \mu, \phi \star \mu', \phi(x), u)),$$



which is well posed by definition of  $\zeta_{\mathbb{R}}$  and Lemma C.2. Now, fix  $\xi \sim \mu$ ,  $U$  a uniform variable independent of  $\xi$ , and define  $\xi' := \zeta(\mu, \mu', \xi, U)$ . By definition of  $\zeta$ , it's clear that  $\xi' \sim \mu'$ , and

$$\mathbb{E}[d(\phi(\xi), \phi(\xi'))] = \mathcal{W}(\phi \star \mu, \phi \star \mu'). \quad (\text{C.2})$$

Fix  $\epsilon > 0$ . We are looking for  $\eta, \delta > 0$  such that

$$\mathcal{W}(\mu, \mu') < \eta \Rightarrow \mathcal{W}(\phi \star \mu, \phi \star \mu') < \delta \Leftrightarrow \mathbb{E}[d(\phi(\xi), \phi(\xi'))] < \delta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon.$$

Let us first show that there exists  $\delta > 0$  such that  $\mathbb{E}[d(\phi(\xi), \phi(\xi'))] < \delta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon$ . Fix  $\gamma > 0$  such that  $d(x, x') < \gamma \Rightarrow d(\phi^{-1}(x), \phi^{-1}(x')) < \frac{\epsilon}{2}$ . Denoting by  $\Delta_{\mathcal{X}}$  the diameter of  $\mathcal{X}$ , we then have

$$\mathbb{E}[d(\xi, \xi')] \leq \mathbb{E}[d(\xi, \xi') \mathbf{1}_{d(\phi(\xi), \phi(\xi')) < \gamma}] + \frac{\Delta_{\mathcal{X}}}{\gamma} \mathbb{E}[d(\phi(\xi), \phi(\xi'))] \leq \frac{\epsilon}{2} + \frac{\Delta_{\mathcal{X}}}{\gamma} \mathbb{E}[d(\phi(\xi), \phi(\xi'))],$$

so that we can choose  $\delta = \frac{\gamma}{\Delta_{\mathcal{X}}} \frac{\epsilon}{2}$ . On the other hand, by uniform continuity of  $\phi$  and by definition of the Wasserstein metric, there exists  $\eta > 0$  such that  $d(\mu, \mu') < \eta \Rightarrow \mathcal{W}(\phi \star \mu, \phi \star \mu') < \delta$ . From (C), we thus conclude that  $d(\mu, \mu') < \eta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon$ .  $\square$

## References

- [1] E. Bayraktar, A. Cosso, and H. Pham. Randomized dynamic programming principle and Feynman-Kac representation for optimal control of McKean-Vlasov dynamics. *Transactions of the American Mathematical Society*, 370:2115–2160, 2018.
- [2] A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*. Springer, 2013.
- [3] D. P. Bertsekas. *Dynamic programming and optimal control, Vol II, approximate dynamic programming*. Athena Scientific, Belmont MA, 2012, 4th edition.
- [4] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 50(2):539–563, 2014.
- [5] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications vol I. and II*. Probability Theory and Stochastic Modelling. Springer, 2018.
- [6] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. arXiv: 1910.12802v1, 2019.
- [7] M.F. Djete. Extended mean field control problem: a propagation of chaos result. *arXiv:2006.12996*, 2020.
- [8] M.F. Djete, D. Possamai, and X. Tan. McKean-Vlasov optimal control: the dynamic programming principle. arXiv:1907.08860, 2019.
- [9] J. Fontbana, H. Guérin, and S. Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electronic communications in probability*, 15:124–133, 2010.
- [10] M. Fornasier, S. Lisini, C. Orrieri, and G. Savaré. Mean-field optimal control as Gamma-limit of finite agent controls. *European Journal of Applied Mathematics*, pages 1–34, 2018.
- [11] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

- [12] H. Gu, X. Guo, X. Wei, and R. Xu. Dynamic programming principles for learning MFCs. arXiv:1911.07314, 2019.
- [13] X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. *arXiv: 1901.09585*, 2019.
- [14] O. Kallenberg. *Foundations of Modern Probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.
- [15] D. Lacker. Limit theory for controlled McKean-Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(3):1641–1672, 2017.
- [16] M. Laurière and O. Pironneau. Dynamic programming for mean-field type controls. *Journal of Optimization Theory*, 169(3):902–924, 2016.
- [17] M. Laurière and L. Tangpi. Convergence of large population games to mean field games with interaction through the controls. *arXiv:2004.08351*, 2020.
- [18] H. Pham and X. Wei. Discrete time McKean-Vlasov control problem: a dynamic programming approach. *Applied Mathematics and Optimization*, 74(3):487–506, 2016.
- [19] H. Pham and X. Wei. Dynamic programming for optimal control of stochastic McKean-Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55:1069–1101, 2017.
- [20] S.T. Rachev and L. Rüschendorf. *Mass transportation problems*. Springer Verlag, 1998.
- [21] R.S. Sutton and A.G. Barto. *Reinforcement learning: an introduction*. Cambridge, MA, 2017, 2nd edition.
- [22] A. Van der Vaart and J.A. Wellner. *Weak convergence of empirical processes*. Springer Verlag, 1996.
- [23] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.