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

Médéric Motte, Huyên Pham

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

Médéric MOTTE * Huyên PHAM †

December 16, 2019

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 ϵ-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

1 Introduction **2**

2 The N-agent and the limiting McKean-Vlasov MDP **5**

3 Lifted MDP on $\mathcal{P}(\mathcal{X})$ **13**
3.1 Case without common noise 14
3.2 Case with discrete state space $\mathcal{X}$ 16

4 General case and Bellman fixed point equation in $\mathcal{P}(\mathcal{X})$ **18**
4.1 A general lifted MDP on $\mathcal{P}(\mathcal{X})$ 19
4.2 Bellman fixed point on $\mathcal{P}(\mathcal{X})$ 21
4.3 Building ϵ-optimal randomized feedback controls 23
4.4 Relaxing the randomization hypothesis 28
4.5 Computing value function and ϵ-optimal strategies in CMKV-MDP 31

5 Conclusion **35**

A Some useful results on conditional law **35**

B Wasserstein convergence of the empirical law **37**

C Proof of coupling results **37**

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 [14], [16], [1], [7]. We point out the work [13], which is the first paper to rigorously connect mean-field control to large systems of controlled processes, see also the recent paper [9], 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, and shortly abbreviated in the sequel as CMKV-MDP. Our set-up is the mathematical framework for a theory of reinforcement learning with mean-field interaction, and is notably motivated by the recent popularization of targeted advertising, in which controls are naturally of open-loop form as the individuals states are inaccessible. The common noise is also a feature of interest to model the impact of public data on the population and to understand how it may affect the strategy of the social planner/influencer.

Compared to continuous-time models, discrete-time McKean-Vlasov control problems have been less studied in the literature. In [15], 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], [11] addressed Bellman equations for MFC problems in the context of reinforcement learning. The paper [11] 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], [11] 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 $\mathbb{L}^1$ 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 discrete 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 ϵ-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], [18].

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. Reinforcement learning algorithms and practical implementation are postponed to a companion paper with applications to model for targeted advertising in social networks.

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 discrete. 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 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 pr_1 (resp. 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 $\Phi : \mathcal{X}_1 \to \mathcal{X}_2$, and a positive measure μ_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

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

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) σ-algebra on $\mathcal{P}(\mathcal{X}_1)$, that is the smallest σ-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 ν 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 μ_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

$$(\mu_1 \cdot \nu)(B_1 \times B_2) = \int_{B_1 \times B_2} \mu_1(\mathrm{d}x_1)\nu(x_1, \mathrm{d}x_2), \quad \forall B_1 \in \Sigma_1,\ B_2 \in \Sigma_2.$$

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

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

where $\boldsymbol{\Pi}(\mu, \mu')$ is the set of probability measures on $\mathcal{Y} \times \mathcal{Y}$ with marginals μ and μ', i.e., $\mathrm{pr}_1 \star \boldsymbol{\mu} = \mu$, and $\mathrm{pr}_2 \star \boldsymbol{\mu} = \mu'$. Since $(\mathcal{Y}, d)$ is compact, it is known (see e.g. Corollary 6.13 in [20]) that the Borel σ-algebra generated by the Wasserstein metric coincides with the cylinder σ-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

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

where $L_{lip}(\mathcal{Y}; R)$ is the set of Lipschitz continuous functions ϕ 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 $\boldsymbol{d}((x, a), (x', a')) = d(x, x') + d_A(a, a')$, $x, x' \in \mathcal{X}$, $a, a' \in A$, and the associated space of probability measure $\mathcal{P}(\mathcal{X} \times A)$, equipped with its Wasserstein distance $\boldsymbol{W}$. Let G, E, and E^0 be three measurable spaces, representing respectively the initial information, idiosyncratic noise, and common noise spaces.

We denote by Π_{OL} the set of sequences $(\pi_t)_{t \in \mathbb{N}}$ (called *open-loop policies*) where π_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

- $(\Gamma^i, \xi^i)_{i\in\mathbb{N}^\star}$ (initial informations and initial states) valued in $G \times \mathcal{X}$
- $(\varepsilon_t^i)_{i\in\mathbb{N}^\star, t\in\mathbb{N}}$ (idiosyncratic noises) valued in E with probability distribution λ_ε
- $\varepsilon^0 := (\varepsilon_t^0)_{t\in\mathbb{N}}$ (common noise) valued in E^0.

Without loss of generality, we may assume that $\mathcal{F}$ contains an atomless random variable, i.e., $\mathcal{F}$ is rich enough, so that any probability measure ν 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 π, we associate an *open-loop control* for individual $i \in \mathbb{N}^*$ as the process $\alpha^{i,\pi}$ defined by

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

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 $i = 1, \ldots, N$ in an N-agent MDP is given by the dynamical system

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

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

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

where the reward f is a mesurable 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 β is a positive discount factor in $[0, 1)$. The influencer's renormalized and expected gains are

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

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 β 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

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

Here, we denote by $\mathbb{P}^0$ and $\mathbb{E}^0$ the conditional probability and expectation knowing the common noise ε^0, and then, given a random variable Y valued in $\mathcal{Y}$, we denote by $\mathbb{P}^0_Y$ or $\mathcal{L}^0(Y)$ its conditional law knowing ε^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

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

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

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

Problem (2.1)-(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*

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

Then, for any $\pi \in \Pi_{OL}$, $X_t^{i,N,\pi} \underset{N\to\infty}{\overset{a.s.}{\to}} X_t^{i,\pi}$, $i \in \mathbb{N}^*$, $t \in \mathbb{N}$, and

$$\boldsymbol{W}\Big(\frac{1}{N}\sum_{i=1}^N \delta_{(X_t^{i,N,\pi},\alpha_t^{i,\pi})}, \mathbb{P}^0_{(X_t^{i,\pi},\alpha_t^{i,\pi})}\Big) \underset{N\to\infty}{\overset{a.s.}{\to}} 0, \qquad \frac{1}{N}\sum_{i=1}^N d(X_t^{i,N,\pi}, X_t^{i,\pi}) \underset{N\to\infty}{\to} 0 \quad a.s.$$

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

$$J_i^{N,\pi} \underset{N\to\infty}{\overset{a.s.}{\longrightarrow}} J_i^\pi, \qquad J^{N,\pi} \underset{N\to\infty}{\overset{a.s.}{\longrightarrow}} J^\pi, \qquad V^{N,\pi} \underset{N\to\infty}{\longrightarrow} V^\pi, \qquad \textit{and} \quad \liminf_{N\to\infty} V^N \;\geq\; V.$$

Proof. Fix $\pi \in \Pi_{OL}$. We omit the dependence of the state processes and control on π, 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 $\nu_t := \mathbb{P}^0_{(X_t^i,\alpha_t^i)}$.
(1) We first prove the convergence of trajectories, for all $i \in \mathbb{N}^\star$,

$$\mathbb{P}\big[\lim_{N\to\infty}(X_t^{i,N}, \nu_t^N) = (X_t^i, \nu_t)\big] = 1, \qquad \mathbb{P}\Big[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N d(X_t^{i,N}, X_t^i) = 0\Big] = 1,$$

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 $N \in \mathbb{N}_\star$ and $i \in \mathbb{N}^\star$, which obviously implies that $X_0^{i,N} \underset{N\to\infty}{\overset{a.s.}{\to}} X_0^i$ and $\frac{1}{N}\sum_{i=1}^N d(X_0^{i,N}, X_0^i) \underset{N\to\infty}{\to} 0$. 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 [19]), and the fact that Wasserstein distance metrizes weak convergence.

- *Hereditary Property.* We have

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

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

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

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

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

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

Let us now show that $\frac{1}{N}\sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i) \overset{a.s.}{\underset{N\to\infty}{\to}} 0$. From (2.3), 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 $D_N := \max[d(X_t^{i,N}, X_t^i), \boldsymbol{W}(\nu_t^N, \nu_t)]$, and

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

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 $\eta > 0$,

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

and thus

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

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

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

As $\eta \to 0$, 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 η 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 ϵ for N large enough.

Let us finally prove that $\boldsymbol{W}(\nu_{t+1}^N, \nu_{t+1}) \overset{a.s.}{\underset{N\to\infty}{\to}} 0$. We have $\boldsymbol{W}(\nu_{t+1}^N, \nu_{t+1}) \le \boldsymbol{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty}) + \boldsymbol{W}(\nu_{t+1}^{N,\infty}, \nu_{t+1})$. To dominate the first term $\boldsymbol{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty})$, notice that, given a variable $U_N \sim \mathcal{U}(\{1, ..., N\})$, the random measure ν_{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 definiton of the Wasserstein distance, $\boldsymbol{W}(\nu_{t+1}^N, \nu_{t+1}^{N,\infty})$ 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 $\alpha_{t+1}^i = \pi_{t+1}(\Gamma^i, (\varepsilon_s^i)_{s\le t+1}, (\varepsilon_s^0)_{s\le t+1})$, and by Proposition A.1, there exists a measurable function $f_{t+1} : \mathcal{X} \times G \times E^{t+1} \times (E^0)^{t+1} \to \mathcal{X}$ such that $X_{t+1}^{i,N} = f_{t+1}(\Gamma^i, (\varepsilon_s^i)_{s\le t+1}, (\varepsilon_s^0)_{s\le t+1}))$. From Proposition B.1, we then deduce that $\boldsymbol{W}(\nu_{t+1}^{N,\infty}, \nu_{t+1})$ 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 $f(X_t^{i,N}, \alpha_t^i, \nu_t^N) \overset{a.s.}{\underset{N\to\infty}{\to}} f(X_t^i, \alpha_t^i, \nu_t)$ for all $t \in \mathbb{N}$. Thus, as f is bounded, we get by dominated convergence that $J_i^{N,\pi} \overset{a.s.}{\underset{N\to\infty}{\to}} J_i^\pi$. Let us now study the convergence of $J^{N,\pi}$ to J^π. We write

$$|J^{N,\pi} - J^\pi| \;\le\; \frac{1}{N}\sum_{i=1}^N |J_i^{N,\pi} - J_i^\pi| + \Big|\frac{1}{N}\sum_{i=1}^N J_i^\pi - J^\pi\Big| \;=:\; S_N^1 + S_N^2.$$

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,

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

By the same argument as above in (1) for showing that $\frac{1}{N}\sum_{i=1}^N d(X_{t+1}^{i,N}, X_{t+1}^i) \overset{a.s.}{\underset{N\to\infty}{\to}} 0$, we prove that $\Delta_N(f)$ tends a.s. to zero as $N \to \infty$. 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 $J^{N,\pi} \overset{a.s.}{\underset{N\to\infty}{\to}} J^\pi$. By dominated convergence, we then also obtain that $V^{N,\pi} = \mathbb{E}[J^N] \overset{a.s.}{\underset{N\to\infty}{\to}} \mathbb{E}[J^\pi] = V^\pi$. Finally, by considering an ϵ-optimal policy π_ϵ for V, we have

$$\liminf_{N\to\infty} V^N \ge \lim_{N\to\infty} V^{N,\pi_\epsilon} \;=\; V^{\pi_\epsilon} \;\ge\; V - \epsilon,$$

which implies, by sending ϵ to zero, that $\liminf_{N\to\infty} V^N \ge V$. □

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.

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

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

($\mathbf{Hf_{lip}}$) There exists $K_f > 0$, such that for all $a \in A$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$,

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

Remark 2.1 We stress the importance of making the Lipschitz assumption for F in *expectation* only. Indeed, most of the practical applications we have in mind concerns discrete spaces $\mathcal{X}$, for which F cannot be, strictly speaking, Lipschitz, since it maps from a continuous space $\mathcal{P}(\mathcal{X} \times A)$ to a discrete space $\mathcal{X}$. However, F can be Lipschitz *in expectation*, e.g. once integrated w.r.t. the idiosyncratic noise, and it is actually a very natural phenomenon. □

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

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

where ν_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 ν. 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* $M_N \underset{N\to\infty}{\to} 0$. *Furthermore,*

- *If* $\mathcal{X} \times A \subset \mathbb{R}^d$ *for some* $d \in \mathbb{N}^\star$, *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* δ *covering the compact metric set* $\mathcal{X} \times A$ *with diameter* $\Delta_{\mathcal{X}\times A}$ *is smaller than* $\mathcal{O}\Big(\big(\frac{\Delta_{\mathcal{X}\times A}}{\delta}\big)^\theta\Big)$ *for* $\theta > 2$, *then* $M_N = \mathcal{O}(N^{-1/\theta})$.

Proof. The first point is proved in [10], and the second one in [4]. □

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

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

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

Furthermore, if we assume ($\mathbf{Hf_{lip}}$), *then for any* $\eta > 0$ *and* $\gamma = \min\big[1, \frac{|\ln\beta|}{(\ln 2K_F)_+} - \eta\big]$, *there exists a constant* $C = C(K_F, K_f, \beta, \gamma)$ *(explicit in the proof) such that for* N *large enough,*

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

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 for* N *large enough, namely* $M_N^\gamma = \mathcal{O}(\epsilon)$.

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 $\nu_t^\pi := \mathbb{P}^0_{(X_t^{i,\pi},\alpha_t^i)}$. By definition, $\alpha_t^{i,\pi} = \pi_t(\Gamma^i, (\varepsilon_s^i)_{s\le t}, (\varepsilon_s^0)_{s\le t})$, and by Lemma A.1, we have $X_t^{i,\pi} = f_t^\pi(\xi^i, \Gamma^i, (\varepsilon_s^i)_{s\le t}, (\varepsilon_s^0)_{s\le t})$ 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 (2.9), we have

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

Let us now prove (2.10) 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, $\mathbb{E}[d(X_{t+1}^{i,N,\pi}, X_{t+1}^{i,\pi})] = \mathbb{E}\big[\Delta\big(X_t^{i,N,\pi}, X_t^{i,\pi}, \alpha_t^i, \nu_t^{N,\pi}, \nu_t^\pi, \varepsilon_{t+1}^0\big)\big]$, where

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

by ($\mathbf{HF_{lip}}$). On the other hand, we have

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{W}(\nu_t^{N,\pi},\nu_t^\pi)\big] &\le \mathbb{E}\big[\boldsymbol{W}(\nu_t^{N,\pi},\nu_t^{N,\infty,\pi})\big] + \mathbb{E}\big[\boldsymbol{W}(\nu_t^{N,\infty,\pi},\nu_t^\pi)\big]\\ &\le \mathbb{E}[d(X_t^{i,N,\pi}, X_t^{i,\pi})] + M_N,\end{aligned} \tag{2.15}$$

where we used the fact that $\boldsymbol{W}(\nu_t^{N,\pi},\nu_t^{N,\infty,\pi}) \le \frac{1}{N}\sum_{i=1}^N d(X_t^{i,N,\pi}, X_t^{i,\pi})$, and (2.13). It follows from (2.14) that

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

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 (2.10). Plugging (2.10) into (2.15) then yields (2.11).

Let us now prove the convergence of gains. From ($\mathbf{Hf_{lip}}$), and (2.15), we have

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

where we set $\delta_t^N := \sup_{\pi\in\Pi_{OL}} \mathbb{E}[d(X_t^{i,N,\pi}, X_t^{i,\pi})]$. From (2.16), we have: $\delta_{t+1}^N \le 2K_F\delta_t^N + K_F M_N$, $t\in\mathbb{N}$, with $\delta_0^N = 0$, and so by induction:

$$\delta_t^N \le \frac{K_F}{1-2K_F}M_N + s_t\big(\frac{K_F}{|2K_F-1|}M_N\big), \quad s_t(m) := m(2K_F)^t,\ m\ge 0.$$

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

$$\sum_{t=0}^\infty \beta^t\delta_t^N \le \frac{K_F}{(1-2K_F)(1-\beta)}M_N + S\big(\frac{K_F}{|2K_F-1|}M_N\big) \tag{2.18}$$

$$S(m) := \sum_{t=0}^\infty \beta^t \min\big[s_t(m); \Delta_{\mathcal{X}}\big], \quad m\ge 0.$$

Since the sequence $(M_N)_N$ converges to zero, we may restrict the study of the function S to the interval $[0, \Delta_{\mathcal{X}}]$, and we notice that it satisfies the dynamic programming relation

$$S(m) = \; m + \beta S\big(\min[s_1(m), \Delta_{\mathcal{X}}]\big), \quad m \in [0, \Delta_{\mathcal{X}}].$$

Therefore, S can be viewed as the unique fixed point in the Banach space $L^\infty([0, \Delta_{\mathcal{X}}])$ of the β-contracting operator $\mathcal{H}$ defined by $[\mathcal{H}\varphi](m) = m + \beta\varphi(\min[s_1(m), \Delta_{\mathcal{X}}])$, and is obtained as the limit of fixed point iterations: $S_{n+1} = \mathcal{H}S_n$, $n \in \mathbb{N}$, $S_0 \equiv 0$. Fix $\gamma \geq 0$ to be determined later, and let us show by induction on $n \in \mathbb{N}$, that $S_n(m) \leq K_{\gamma,n} m^\gamma$, $m \in [0, \Delta_{\mathcal{X}}]$, for some suitable sequence $(K_{\gamma,n})_n$. By writing that $S_{n+1} = \mathcal{H}S_n$ at step $n+1$, and using the induction hypothesis, we have for all $m \in [0, \Delta_{\mathcal{X}}]$,

$$S_{n+1}(m) \;\; \leq \;\; m + \beta K_{\gamma,n} s_1(m)^\gamma \;\leq\; \big(\Delta_{\mathcal{X}}^{1-\gamma} + \beta K_{\gamma,n}(2K_F)^\gamma\big) m^\gamma.$$

By setting $K_{\gamma,n+1} := \Delta_{\mathcal{X}}^{1-\gamma} + \beta(2K_F)^\gamma K_{\gamma,n}$, starting with $K_{\gamma,0} = 0$, we see that the required relation is satisfied at iteration $n+1$. Now, by taking $\gamma < \frac{|\ln(\beta)|}{(\ln 2K_F)_+}$, we have $K_{\gamma,n} = \Delta_{\mathcal{X}}^{1-\gamma} \frac{1-(\beta(2K_F)^\gamma)^n}{1-\beta(2K_F)^\gamma} \to K_\gamma := \frac{\Delta_{\mathcal{X}}^{1-\gamma}}{1-\beta(2K_F)^\gamma}$, as n goes to infinity, and thus

$$S(m) \leq \; K_\gamma m^\gamma, \quad \forall m \in [0, \Delta_{\mathcal{X}}].$$

From (2.17)-(2.18), it follows that for N large enough (so that $M_N < \frac{\Delta_{\mathcal{X}}|2K_F - 1|}{K_F}$), and taking $\gamma = \min\big[1, \frac{|\ln\beta|}{(\ln 2K_F)_+} - \eta\big]$, for $\eta > 0$,

$$|V^{N,\pi} - V^\pi| \leq C M_N^\gamma, \quad \pi \in \Pi_{OL},$$

for some constant C that is explicit in terms of K_f, K_F, β and γ. This concludes the proof. □

Remark 2.2 If the Lipschitz constant in ($\mathbf{HF_{lip}}$) satisfies $\beta 2K_F < 1$, then we can take $\gamma = 1$ in the rate of convergence (2.12) of the optimal value, and this can be directly derived from (2.17) since in this case $\sum_{t=0}^\infty (\beta 2K_F)^t$ is finite and so $\sum_{t=0}^\infty \beta^t \delta_t^N = \mathcal{O}(M_N)$. The function S in the above proof is introduced for dealing with the case when $\beta 2K_F > 1$.

In 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

$$\tilde{M}_N := \max\Big\{ \sup_{\mu \in \mathcal{P}(\mathcal{X})} \mathbb{E}[\mathcal{W}(\mu_N, \mu)], \; \sup_{\upsilon \in \mathcal{P}(A)} \mathbb{E}[\mathcal{W}_A(\upsilon_N, \upsilon)] \Big\},$$

instead of M_N in (2.9), where here μ_N (resp. υ_N) is the empirical measure associated to μ (resp. υ) $\in \mathcal{P}(\mathcal{X})$ (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 $M_N = \mathcal{O}\big(N^{-1/2}\log(1+N)\big)$. □

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 (2.2), 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 Γ, initial state by ξ, idiosyncratic noise by $(\varepsilon_t)_{t\in\mathbb{N}}$. In the sequel, we shall denote by $\mathcal{G} = \sigma(\Gamma)$ the σ-algebra generated by the random variable Γ, 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 α, which is adapted to the filtration generated by $(\Gamma, (\varepsilon_s)_{s\le t}, (\varepsilon^0_s)_{s\le t})_{t\in\mathbb{N}}$, and associated to an open-loop policy by: $\alpha_t = \alpha^\pi_t := \pi_t(\Gamma, (\varepsilon_s)_{s\le t}, (\varepsilon^0_s)_{s\le t})$ 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

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

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

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

where we stress the dependence upon the initial state ξ. 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 ε^0. Given an open-loop control $\alpha \in \mathcal{A}$, and its state process $X = X^{\xi,\alpha}$, denote by $\{\mu_t = \mathbb{P}^0_{X_t}, t \in \mathbb{N}\}$, the random $\mathcal{P}(\mathcal{X})$-valued process, and notice from Proposition A.1 that $(\mu_t)_t$ is $\mathbb{F}^0$-adapted. From (3.1), and recalling the pushforward measure notation, we have

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

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

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

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 $\mathbb{P}^0_{(X_t,\alpha_t)} = \mu_t \cdot \hat{\alpha}_t$. The dynamics relation (3.2) is then written as

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

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

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

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

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

with the conditional expectation term equal to

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

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

$$\hat{f}(\mu, \hat{a}) = \int_{\mathcal{X}\times A} f(x, a, \mu \cdot \hat{a})(\mu \cdot \hat{a})(\mathrm{d}x, \mathrm{d}a). \tag{3.4}$$

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 (3.3), a discount factor $\beta \in [0, 1)$, and a reward function $\hat{f}$ as in (3.4). A key point is to endow $\hat{A}(\mathcal{X})$ with a suitable σ-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

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

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

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

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

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

where the bounded function $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \to \mathbb{R}$ is defined as in (3.4). 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 (3.2)-(3.4) (by taking directly law under $\mathbb{P}$ instead of $\mathbb{P}^0$), we clearly see that for any $\alpha \in \mathcal{A}$, $V^\alpha(\xi) = \widehat{V}^{\hat{\alpha}}(\mu)$, with $\mu = \mathcal{L}(\xi)$, and $\hat{\alpha} = \mathcal{R}_\xi(\alpha)$ where $\mathcal{R}_\xi$ is the relaxed operator

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

It follows that $V(\xi) \leq \widehat{V}(\mu)$. 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 σ-algebra generated by ξ is equal to $\mathcal{G}$, then for any $\alpha \in \mathcal{A}$, α_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 $\kappa_0 = \hat{a} \in \hat{A}(\mathcal{X})$. 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 Γ 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 ξ, so that one can freely randomize via the uniform random variable U the first action given ξ according to any probability kernel $\hat{a}$. Moreover, one can extract from U, by standard separation of the decimals of U (see Lemma 2.21 in [12]), an i.i.d. sequence of uniform variables $(U_t)_{t\in\mathbb{N}}$, which are $\mathcal{G}$-measurable, independent of ξ, and can then be used to randomize the subsequent actions. □

Theorem 3.1 (Correspondence in the no common noise case)
Assume that **Rand**(ξ, Γ) *holds true. Then $\mathcal{R}_\xi$ is surjective from $\mathcal{A}$ into $\widehat{\mathcal{A}}$, and we have $V(\xi) = \widehat{V}(\mu)$, for $\mu = \mathcal{L}(\xi)$. Moreover, if $\alpha^\epsilon \in \mathcal{A}$ is an ϵ-optimal control for $V(\xi)$, then $\mathcal{R}_\xi(\alpha^\epsilon) \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$, and conversely, if $\hat{\alpha}^\epsilon \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$, then any $\alpha^\epsilon \in \mathcal{R}_\xi^{-1}(\hat{\alpha}^\epsilon)$ is an ϵ-optimal control for $V(\xi)$.*

Proof. In view of the above discussion, we only need to prove the surjectivity of $\mathcal{R}_\xi$. Fix a control $\kappa \in \widehat{\mathcal{A}}$ for the MDP on $\mathcal{P}(\mathcal{X})$. By Proposition A.3, for all $t \in \mathbb{N}$, there exists a measurable function $\mathrm{a}_t : \mathcal{X} \times [0,1] \to A$ such that $\mathbb{P}_{\mathrm{a}_t(x,U)} = \kappa_t(x)$, for all $x \in \mathcal{X}$. It is then clear that the control α defined recursively by $\alpha_t := \mathrm{a}_t(X_t^{\xi,\alpha}, U_t)$, where $(U_t)_t$ is an i.i.d. sequence of $\mathcal{G}$-measurable uniform variables independent of ξ under **Rand**(ξ, Γ), satisfies $\mathcal{L}(\alpha_t \mid X^{\xi,\alpha} = x) = \kappa_t(x)$ (observing that U_t is independent of $X_t^{\xi,\alpha}$), and thus $\hat{\alpha} = \kappa$, which proves the surjectivity of $\mathcal{R}_\xi$. □

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 ξ only via its probability law $\mu = \mathcal{L}(\xi)$, for ξ satisfying the randomization hypothesis. □

3.2 Case with discrete state space $\mathcal{X}$

We consider the case with common noise but when the state space $\mathcal{X}$ is discrete, 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 σ-algebra of the Wasserstein metric space $\mathcal{P}(A)$.

Lemma 3.1 *Suppose that $\#\mathcal{X} = n < \infty$. Then, $\hat{F}$ in (3.3) 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 (3.4) 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 Corollary 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 discrete case, $\mu \cdot \hat{a}$ is here simply equal to $\sum_{x\in\mathcal{X}} \mu(x)\hat{a}(x)$ and, thus Ψ is clearly measurable. □

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 $\widehat{\mathcal{A}}$ the set of $\mathbb{F}^0$-adapted processes valued in $\mathcal{P}(A)^n$, and given $\kappa \in \widehat{\mathcal{A}}$, consider the controlled dynamics in $\mathbb{S}^{n-1}$

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

the associated expected gain and value function

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

We aim to prove the correspondence and equivalence between the CMKV-MDP and the MDP (3.5)-(3.6). From the derivation in (3.2)-(3.4) and by Lemma 3.1, we see that for any $\alpha \in \mathcal{A}$, $V^\alpha(\xi) = \widehat{V}^{\hat{\alpha}}(\mu)$, where $\mu = \mathcal{L}(\xi)$, and $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ where $\mathcal{R}^0_\xi$ is the relaxed operator

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

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

Lemma 3.2 *Assume that Γ 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* **Rand**(ξ, Γ).

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 Ω with this probability, Γ 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 ξ. □

Theorem 3.2 (Correspondance with the MDP on $\mathcal{P}(\mathcal{X})$ in the $\mathcal{X}$ discrete case) *Assume that $\mathcal{G}$ is rich enough. Then $\mathcal{R}^0_\xi$ is surjective from $\mathcal{A}$ into $\widehat{\mathcal{A}}$, and $V(\xi) = \widehat{V}(\mu)$, for any $\mu \in \mathcal{P}(\mathcal{X})$, $\xi \in L^0(\mathcal{G};\mathcal{X})$ s.t. $\mu = \mathcal{L}(\xi)$. Moreover, if $\alpha^\epsilon \in \mathcal{A}$ is an ϵ-optimal control for $V(\xi)$, then $\mathcal{R}^0_\xi(\alpha^\epsilon) \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$. Conversely, if $\hat{\alpha}^\epsilon \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$, then any $\alpha^\epsilon \in (\mathcal{R}^0_\xi)^{-1}(\hat{\alpha}^\epsilon)$ is an ϵ-optimal control for $V(\xi)$.*

Proof. From the derivation in (3.5)-(3.7), we only need to prove the surjectivity of $\mathcal{R}^0_\xi$. Fix $\kappa \in \widehat{\mathcal{A}}$ and let $\boldsymbol{\pi}_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$ be such that $\kappa_t = \boldsymbol{\pi}_t((\varepsilon^0_s)_{s\le t})$. As $\mathcal{X}$ is discrete, by definition of the σ-algebra on $\hat{A}(\mathcal{X})$, $\boldsymbol{\pi}_t$ 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 Corollary A.1, we know that $\phi \star \boldsymbol{\pi}_t$ 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^{-1}_m the generalized inverse of its distribution function, and it is known that the mapping $m \in (\mathcal{P}(\mathbb{R}), \mathcal{W}) \mapsto F^{-1}_m \in (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1)$ is an isometry and is thus measurable. Therefore, $F^{-1}_{\phi\star\boldsymbol{\pi}_t}$ 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 $n \to \infty$. Therefore, the mapping

$$\begin{aligned} \mathrm{a}_t : (E^0)^t \times \mathcal{X} \times [0,1] &\longrightarrow A \\ ((e^0_s)_{s\le t}, x, u) &\longmapsto \phi^{-1} \circ F^{-1}_{\phi\star\boldsymbol{\pi}_t((e^0_s)_{s\le t}, x)}(u) \end{aligned}$$

is measurable. We thus define, by induction, $\alpha_t := \mathrm{a}_t((\varepsilon^0_s)_{s\le t}, X^{\xi,\alpha}_t, U_t)$. By construction and by the generalized inverse simulation method, it is clear that $\hat{\alpha}_t = \kappa_t$. □

Remark 3.3 We point out that when both state space $\mathcal{X}$ and action space A are discrete, equipped with the metrics $d(x,x') := \mathbf{1}_{x\neq y}$, $x, x' \in \mathcal{X}$ 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 ($\mathbf{HF_{lip}}$) and ($\mathbf{Hf_{lip}}$) 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

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

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. □

Remark 3.4 As in the no common noise case, the correspondence result in the $\mathcal{X}$ discrete case 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 σ-algebra for the lifted MDP to be well-posed: on one hand, this σ-algebra has to be large enough to make the functions $\hat{F} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \times E^0 \to \mathcal{P}(\mathcal{X})$ and $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \to \mathbb{R}$ measurable, and on the other hand, it should be small enough to make the process $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ $\mathbb{F}^0$-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 $\mathcal{R}^0_\xi$ from $\mathcal{A}$ into $\widehat{\mathcal{A}}$. 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 σ-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 $\boldsymbol{\pi}_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$ as an element of $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$, which is crucial because the control α (such that $\hat{\alpha} = \kappa$) is defined with α_t explicitly depending upon $\boldsymbol{\pi}_t((\varepsilon^0_s)_{s\le t}, X_t)$.

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 $\mathcal{P}(\mathcal{X})$. □

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 ϵ-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: $V(\xi) \le \tilde{V}(\mu)$, 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 $V^\star$.

(iii) Under the randomization hypothesis for ξ, we show the existence of an ϵ-randomized feedback policy, which yields both an ϵ-randomized feedback control for the CMKV-MDP and an ϵ-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 $\tilde{V} = V^\star$, 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 (3.2) describing the evolution of $\mu_t = \mathbb{P}^0_{X_t}$, $t \in \mathbb{N}$, for a state process $X_t = X_t^{\xi,\alpha}$ controlled by $\alpha \in \mathcal{A}$:

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

Now, instead of decoupling as in Section 3, the conditional law of the pair (X_t, α_t), as $\mathbb{P}^0_{(X_t,\alpha_t)} = \mu_t \cdot \hat{\alpha}_t$ where $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ is the relaxed control in (3.7), we directly consider the control process $\boldsymbol{\alpha}_t = \mathbb{P}^0_{(X_t,\alpha_t)}$, $t \in \mathbb{N}$, which is $\mathbb{F}^0$-adapted (see Proposition A.1), and valued in the space of probability measures $\boldsymbol{A} := \mathcal{P}(\mathcal{X} \times A)$, naturally endowed with the σ-algebra of its Wasserstein metric. Notice that this $\boldsymbol{A}$-valued control $\boldsymbol{\alpha}$ obtained from the CMKV-MDP has to satisfy by definition the marginal constraint $\mathrm{pr}_1 \star \boldsymbol{\alpha}_t = \mu_t$ 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 ξ.*
- *(i) When $\mathcal{X} \subset \mathbb{R}$:*

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

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

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

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 μ, $\mu' \in \mathcal{P}(\mathcal{X})$, there exists ξ and ξ' random variables with $\mathcal{L}(\xi) = \mu$, $\mathcal{L}(\xi') = \mu'$ such that $\mathcal{W}(\mu,\mu') = \mathbb{E}\big[d(\xi,\xi')\big]$. A similar measurable optimal coupling is proved in [8] under the assumption that there exists a transfer function realizing an optimal coupling between μ and μ'. However, such transfer function does not always exist, for instance when μ has atoms but not μ'. Lemma 4.1 builds a measurable coupling without making such assumption (essentially using the uniform variable U to randomize when μ has atoms). □

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

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

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

Lemma 4.2 (Measurable coupling projection)
The coupling projection $\boldsymbol{p}$ is a measurable function from $\mathcal{P}(\mathcal{X}) \times \boldsymbol{A}$ into $\boldsymbol{A}$, and for all $(\mu, \mathbf{a}) \in \mathcal{P}(\mathcal{X}) \times \boldsymbol{A}$:

$$pr_{_1} \star \boldsymbol{p}(\mu, \mathbf{a}) \;=\; \mu, \quad \textit{and if} \quad pr_1 \star \mathbf{a} = \mu, \;\; \textit{then } \boldsymbol{p}(\mu, \mathbf{a}) = \mathbf{a}.$$

Proof. The only result that is not trivial is the measurability of $\boldsymbol{p}$. Observe that $\boldsymbol{p}(\mu, \mathbf{a}) = g(\mu, \mathbf{a}, \cdot, \cdot, \cdot) \star (\mathbf{a} \otimes \mathcal{U}([0,1]))$ where g is the measurable function

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

We thus conclude by Corollary A.1. □

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

$$\mu_{t+1} = \; \tilde{F}(\mu_t, \boldsymbol{\alpha}_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N}, \tag{4.2}$$

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

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

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

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

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

$$\mu_{t+1} = \; \tilde{F}(\mu_t, \boldsymbol{\nu}_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N}, \; \mu_0 = \mu \in \mathcal{P}(\mathcal{X}), \tag{4.3}$$

with associated expected gain/value function

$$\widetilde{V}^{\boldsymbol{\nu}}(\mu) = \; \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \tilde{f}(\mu_t, \boldsymbol{\nu}_t)\Big], \qquad \widetilde{V}(\mu) \;=\; \sup_{\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}} \widetilde{V}^{\boldsymbol{\nu}}(\mu). \tag{4.4}$$

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

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

By construction from (4.2), we see that $\mu_t = \mathbb{P}^0_{X^{\xi,\alpha}_t}$, $t \in \mathbb{N}$, follows the dynamics (4.3) with the control $\boldsymbol{\nu} = \boldsymbol{\mathcal{L}}^0_\xi(\alpha) \in \boldsymbol{\mathcal{A}}$. 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

$$\begin{aligned} V^\alpha(\xi) = \; & \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \mathbb{E}^0\big[f(X^{\xi,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\xi,\alpha}_t, \alpha_t)})\big]\Big] \\ = \; & \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \tilde{f}(\mathbb{P}^0_{X^{\xi,\alpha}_t}, \boldsymbol{\alpha}_t)\Big] \;=\; \widetilde{V}^{\boldsymbol{\alpha}}(\mu), \quad \text{with } \; \mu = \mathcal{L}(\xi). \end{aligned} \tag{4.5}$$

It follows that $V(\xi) \leq \widetilde{V}(\mu)$, 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 (4.3)-(4.4) on $\mathcal{P}(\mathcal{X})$. By defining this MDP on the canonical space $(E^0)^{\mathbb{N}}$, an open-loop control $\boldsymbol{\nu} = (\boldsymbol{\nu}_t)_t \in \boldsymbol{\mathcal{A}}$ is identified with the open-loop policy $\boldsymbol{\nu} = (\nu_t)_t$ where ν_t is a measurable function from $(E^0)^t$ into $\boldsymbol{A}$ via the relation $\boldsymbol{\nu}_t = \nu_t(\varepsilon_1^0, \ldots, \varepsilon_t^0)$, with the convention that $\boldsymbol{\nu}_0 = \nu_0$ is simply a constant in $\boldsymbol{A}$. Given $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, and $e^0 \in E^0$, we denote by $\vec{\boldsymbol{\nu}}(e^0) = (\vec{\nu}_t(e^0))_t \in \boldsymbol{\mathcal{A}}$ the shifted open-loop policy defined by $\vec{\nu}_t(e^0)(.) = \nu_{t+1}(e^0, .)$, $t \in \mathbb{N}$. Given $\mu \in \mathcal{P}(\mathcal{X})$, and $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we denote by $(\mu_t^{\mu,\boldsymbol{\nu}})_t$ the solution to (4.3), which satisfies the flow property

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

where $\equiv$ means equality in law under $\mathbb{P}$. This implies that the expected gain of this MDP in (4.4) satisfies the relation

$$\widetilde{V}^{\boldsymbol{\nu}}(\mu) = \tilde{f}(\mu, \boldsymbol{\nu}_0) + \beta \mathbb{E}\Big[\widetilde{V}^{\vec{\boldsymbol{\nu}}(\varepsilon_1^0)}(\mu_1^{\mu,\boldsymbol{\nu}})\Big]$$

Recalling that $\mu_1^{\mu,\boldsymbol{\nu}} = \tilde{F}(\mu, \boldsymbol{\nu}_0, \varepsilon_1^0)$, and by definition of $\tilde{V}$ as the value function, we deduce that

$$\widetilde{V}(\mu) \leq \big[\mathcal{T}\widetilde{V}\big](\mu), \tag{4.6}$$

where $\mathcal{T}$ is the Bellman operator on $L^\infty(\mathcal{P}(\mathcal{X}))$, the set of bounded real-valued functions on $\mathcal{P}(\mathcal{X})$, defined for any $W \in L^\infty(\mathcal{P}(\mathcal{X}))$ by

$$[\mathcal{T}W](\mu) := \sup_{\mathbf{a} \in \boldsymbol{A}} \Big\{\tilde{f}(\mu, \mathbf{a}) + \beta \mathbb{E}\big[W\big(\tilde{F}(\mu, \mathbf{a}, \varepsilon_1^0)\big)\big]\Big\}, \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{4.7}$$

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 ε_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 $W \in L^\infty(\mathcal{P}(\mathcal{X}))$, and $\mu \in \mathcal{P}(\mathcal{X})$, we have*

$$[\mathcal{T}W](\mu) = \sup_{\hat{a} \in \hat{A}(\mathcal{X})} [\hat{\mathcal{T}}^{\hat{a}}W](\mu) = \sup_{\mathrm{a} \in L^0(\mathcal{X} \times [0,1]; A)} [\mathbb{T}^{\mathrm{a}}W](\mu), \tag{4.8}$$

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

$$\begin{aligned}
[\hat{\mathcal{T}}^{\hat{a}}W](\mu) &= \hat{f}(\mu, \hat{a}) + \beta \mathbb{E}\big[W\big(\hat{F}(\mu, \hat{a}, \varepsilon_1^0)\big)\big], \\
[\mathbb{T}^{\mathrm{a}}W](\mu) &= \mathbb{E}\Big[f(\xi, \mathrm{a}(\xi, U), \mathcal{L}(\xi, \mathrm{a}(\xi, U))) + \beta W\big(\mathbb{P}^0_{F(\xi, \mathrm{a}(\xi,U), \mathcal{L}(\xi, \mathrm{a}(\xi,U)), \varepsilon_1, \varepsilon_1^0)}\big)\Big],
\end{aligned} \tag{4.9}$$

for any $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$ (it is clear that the right-hand side in (4.9) does not depend on the choice of such (ξ, U)). Moreover, we have

$$[\mathcal{T}W](\mu) = \sup_{\alpha_0 \in L^0(\Omega; A)} \mathbb{E}\Big[f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)) + \beta W\big(\mathbb{P}^0_{F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1, \varepsilon_1^0)}\big)\Big]. \tag{4.10}$$

Proof. Fix $W \in L^\infty(\mathcal{P}(\mathcal{X}))$, and $\mu \in \mathcal{P}(\mathcal{X})$. Let $\mathbf{a}$ be arbitrary in $\boldsymbol{A}$. Since $\boldsymbol{p}(\mu, \mathbf{a})$ has first marginal equal to μ, there exists by Corollary A.2 a probability kernel $\hat{a} \in \hat{A}(\mathcal{X})$ such that $\boldsymbol{p}(\mu, \mathbf{a}) = \mu \cdot \hat{a}$. 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

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

and notice that it is surjective from $L^0(\mathcal{X} \times [0,1]; A)$ into $\hat{A}(\mathcal{X})$, by Lemma A.3. By noting that for any $\mathrm{a} \in L^0(\mathcal{X} \times [0,1]; A)$, and $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$, we have $\mathcal{L}\big(\xi, \mathrm{a}(\xi, U)\big) = \mu \cdot \mathcal{R}(\mathrm{a})$, it follows that $[\mathbb{T}^{\mathrm{a}} W](\mu) = [\hat{\mathcal{T}}^{\mathcal{R}(\mathrm{a})} W](\mu)$. Since $\mathcal{R}$ is surjective, this yields $\mathbb{T}^1 = \sup_{\mathrm{a} \in L^0(\mathcal{X} \times [0,1]; A)} [\mathbb{T}^{\mathrm{a}} W](\mu) =: \mathbb{T}^2$.

Denote by $\mathbb{T}^3$ the right-hand-side in (4.10). 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 μ. Thus, $\boldsymbol{p}(\mu, \mathbf{a}) = \mathcal{L}(\xi, \alpha_0)$, and so

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

We deduce that $\mathbb{T}^3 \leq [\mathcal{T}W](\mu)$, which gives finally the equalities (4.8) and (4.10). □

By standard and elementary arguments, we state the basic properties of the Bellman operator $\mathcal{T}$.

Lemma 4.3 *(i) The operator $\mathcal{T}$ is contracting on $L^\infty(\mathcal{P}(\mathcal{X}))$ with Lipschitz factor β, and admits a unique fixed point in $L^\infty(\mathcal{P}(\mathcal{X}))$, denoted by $V^\star$, hence solution to:*

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

(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 $\mathcal{T}W_1 \leq \mathcal{T}W_2$.

As a consequence of Lemma 4.3, we can easily show the following relation between the value function $\tilde{V}$ of the general lifted MDP, and the fixed point $V^\star$ of the Bellman operator.

Lemma 4.4 *For all $\mu \in \mathcal{P}(\mathcal{X})$, we have*

$$\tilde{V}(\mu) \leq V^\star(\mu).$$

Proof. From (4.6), we have: $\tilde{V} \leq \mathcal{T}\tilde{V} =: V_1$. By iterating this inequality with the operator $\mathcal{T}$, and using the monotonic increasing property of $\mathcal{T}$, we get $\tilde{V} \leq \mathcal{T}^n \tilde{V} =: V_n$. Since the fixed point $V^\star$ of the contracting operator $\mathcal{T}$ is the limit of V_n, as n goes to infinity, this proves the second inequality $\tilde{V} \leq V^\star$. □

4.3 Building ϵ-optimal randomized feedback controls

We aim to prove rigorously the equality $\tilde{V} = V^\star$, 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 ϵ-optimal control for $\tilde{V}$. Notice that it cannot be obtained directly from classical theory of MDP as we consider here open-loop controls $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$ 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 $\boldsymbol{A}$, and in connection with the Bellman operator in (4.7), we introduce, for $\boldsymbol{\pi} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ (the set of measurable functions from $\mathcal{P}(\mathcal{X})$ into $\boldsymbol{A}$) called (measurable) feedback policy, the so-called $\boldsymbol{\pi}$-Bellman operator $\mathcal{T}^{\boldsymbol{\pi}}$ on $L^\infty(\mathcal{P}(\mathcal{X}))$, defined for $W \in L^\infty(\mathcal{P}(\mathcal{X}))$ by

$$[\mathcal{T}^{\boldsymbol{\pi}} W](\mu) = \tilde{f}(\mu, \boldsymbol{\pi}(\mu)) + \beta \mathbb{E}\big[W\big(\tilde{F}(\mu, \boldsymbol{\pi}(\mu), \varepsilon_1^0)\big)\big], \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{4.11}$$

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

Lemma 4.5 *Fix $\boldsymbol{\pi} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$.*

(i) The operator $\mathcal{T}^{\boldsymbol{\pi}}$ is contracting on $L^\infty(\mathcal{P}(\mathcal{X}))$ with Lipschitz factor β, and admits a unique fixed point denoted $\tilde{V}^{\boldsymbol{\pi}}$.

(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 $\mathcal{T}^{\boldsymbol{\pi}} W_1 \leq \mathcal{T}^{\boldsymbol{\pi}} W_2$.

Remark 4.2 It is well-known from MDP theory that the fixed point $\tilde{V}^{\boldsymbol{\pi}}$ to the operator $\mathcal{T}^{\boldsymbol{\pi}}$ is equal to

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

where (μ_t) is the MDP in (4.3) with the feedback and stationary control $\boldsymbol{\nu}^{\boldsymbol{\pi}} = (\boldsymbol{\nu}_t^{\boldsymbol{\pi}})_t \in \boldsymbol{\mathcal{A}}$ defined by $\boldsymbol{\nu}_t^{\boldsymbol{\pi}} = \boldsymbol{\pi}(\mu_t)$, $t \in \mathbb{N}$. In the sequel, we shall then identify by misuse of notation $\tilde{V}^{\boldsymbol{\pi}}$ and $\tilde{V}^{\boldsymbol{\nu}^{\boldsymbol{\pi}}}$ as defined in (4.4). □

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

Remark 4.3 Given $\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, denote by $\boldsymbol{\pi}^{\mathfrak{a}} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ the associated lifted randomized feedback policy, i.e., $\boldsymbol{\pi}^{\mathfrak{a}}(\mu) = \mathcal{L}\big(\xi, \mathfrak{a}(\mu, \xi, U)\big)$, for $\mu \in \mathcal{P}(\mathcal{X})$, and $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$. By definition of the Bellman operator $\mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}}$ in (4.11), and observing that $\boldsymbol{p}(\mu, \boldsymbol{\pi}^{\mathfrak{a}}(\mu)) = \boldsymbol{\pi}^{\mathfrak{a}}(\mu) = \mathcal{L}\big(\xi, \mathrm{a}^\mu(\xi, U)\big)$, where we set $\mathrm{a}^\mu = \mathfrak{a}(\mu, \cdot, \cdot) \in L^0(\mathcal{X} \times [0,1] : A)$, we see (recalling the notation in (4.9)) that for all $W \in L^\infty(\mathcal{P}(\mathcal{X}))$,

$$[\mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}} W](\mu) = [\mathbb{T}^{\mathrm{a}^\mu} W](\mu), \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{4.12}$$

On the other hand, let $\xi \in L^0(\mathcal{G};\mathcal{X})$ be some initial state satisfying the randomization hypothesis $\mathbf{Rand}(\xi,\mathcal{G})$, and denote by $\alpha^{\mathfrak{a}} \in \mathcal{A}$ the randomized feedback stationary control defined by $\alpha_t^{\mathfrak{a}} = \mathfrak{a}(\mathbb{P}^0_{X_t}, X_t, U_t)$, where $X = X^{\xi,\alpha^{\mathfrak{a}}}$ is the state process in (3.1) of the CMKV-MDP, and $(U_t)_t$ is an i.i.d. sequence of uniform $\mathcal{G}$-measurable random variables independent of ξ. By construction, the associated lifted control $\boldsymbol{\alpha}^{\mathfrak{a}} = \boldsymbol{\mathcal{L}}^0_\xi(\alpha^{\mathfrak{a}})$ satisfies $\boldsymbol{\alpha}^{\mathfrak{a}}_{\boldsymbol{t}} = \mathbb{P}^0_{(X_t,\alpha_t^{\mathfrak{a}})} = \boldsymbol{\pi}^{\mathfrak{a}}(\mu_t)$, where $\mu_t = \mathbb{P}^0_{X_t}$, $t \in \mathbb{N}$. Denoting by $V^{\mathfrak{a}} := V^{\alpha^{\mathfrak{a}}}$ the associated expected gain of the CMKV-MDP, and recalling Remark 4.2, we see from (4.5) that $V^{\mathfrak{a}}(\xi) = \tilde{V}^{\boldsymbol{\nu}^{\boldsymbol{\pi}^{\mathfrak{a}}}}(\mu) = \tilde{V}^{\boldsymbol{\pi}^{\mathfrak{a}}}(\mu)$, where $\mu = \mathcal{L}(\xi)$. □

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 4.2 (Verification result)
Fix $\epsilon \geq 0$, and suppose that there exists an ϵ-optimal feedback policy $\boldsymbol{\pi}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ for $V^\star$ in the sense that

$$V^\star \leq \mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star + \epsilon.$$

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

Proof. Since $\tilde{V}^{\boldsymbol{\pi}_\epsilon} = \mathcal{T}^{\boldsymbol{\pi}_\epsilon}\tilde{V}^{\boldsymbol{\pi}_\epsilon}$, and recalling from Lemma 4.4 that $V^\star \geq \tilde{V} \geq \tilde{V}^{\boldsymbol{\pi}_\epsilon}$, we have for all $\mu \in \mathcal{P}(\mathcal{X})$,

$$\left|(V^\star - \tilde{V}^{\boldsymbol{\pi}_\epsilon})(\mu)\right| \leq \left|\mathcal{T}^{\boldsymbol{\pi}_\epsilon}(V^\star - \tilde{V}^{\boldsymbol{\pi}_\epsilon})(\mu) + \epsilon\right| \leq \beta\|V^\star - \tilde{V}^{\boldsymbol{\pi}_\epsilon}\| + \epsilon,$$

where we used the β-contraction property of $\mathcal{T}^{\boldsymbol{\pi}_\epsilon}$ in Lemma 4.5. We deduce that $\|V^\star - \tilde{V}^{\boldsymbol{\pi}_\epsilon}\| \leq \frac{\epsilon}{1-\beta}$, and then, $\tilde{V} \geq \tilde{V}^{\boldsymbol{\pi}_\epsilon} \geq V^\star - \frac{\epsilon}{1-\beta}$, which combined with $V^\star \geq \tilde{V}$, shows the first assertion. Moreover, if $\boldsymbol{\pi}_\varepsilon = \boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$ is a lifted randomized feedback policy, then by Remark 4.3, and under $\mathbf{Rand}(\xi,\mathcal{G})$, we have $V^{\mathfrak{a}_\epsilon}(\xi) = \tilde{V}^{\boldsymbol{\pi}_\epsilon}(\mu)$. Recalling that $V(\xi) \leq \tilde{V}(\mu)$, and together with the first assertion, this proves the required result. □

Remark 4.4 If we can find for any $\epsilon > 0$, an ϵ-optimal lifted randomized feedback policy for $V^\star$, then according to Proposition 4.2, and under $\mathbf{Rand}(\xi,\mathcal{G})$, one could restrict to randomized feedback policies in the computation of the optimal value $V(\xi)$ of the CMKV-MDP, i.e., $V(\xi) = \sup_{\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1];A)} V^{\mathfrak{a}}(\xi)$. Moreover, this would prove that $V(\xi) = \tilde{V}(\mu) = V^\star(\mu)$, 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 $V^\star$ to the Bellman equation, and show via a verification result that V is indeed equal to $V^\star$, hence satisfies the Bellman equation.

By the formulation (4.8) of the Bellman operator in Proposition 4.1, and the fixed point equation satisfied by $V^\star$, we know that for all $\epsilon > 0$, and $\mu \in \mathcal{P}(\mathcal{X})$, there exists $\mathfrak{a}^\mu_\epsilon \in L^0(\mathcal{X} \times [0,1]; A)$ such that

$$V^\star(\mu) \leq [\mathbb{T}^{\mathfrak{a}^\mu_\epsilon} V^\star](\mu) + \epsilon. \tag{4.13}$$

The crucial issue is to prove that the mapping $(\mu, x, u) \mapsto \mathfrak{a}_\epsilon(\mu, x, u) := \mathrm{a}^\mu_\epsilon(x, u)$ is measurable so that it defines a randomized feedback policy $\mathfrak{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, and an associated lifted randomized feedback policy $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$. Recalling the relation (4.12), this would then show that $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$ is a ϵ-optimal lifted randomized feedback policy for $V^\star$, and we could apply the verification result. □

We now address the measurability issue for proving the existence of an ϵ-optimal randomized feedback policy for $V^\star$. The basic idea is to construct as in (4.13) an ϵ-optimal $\mathrm{a}^\mu_\epsilon \in L^0(\mathcal{X} \times [0,1]; A)$ for $V^\star(\mu)$ when μ lies in a suitable finite grid of $\mathcal{P}(\mathcal{X})$, and then "patchs" things together to obtain an ε-optimal randomized feedback policy. This is made possible under some uniform continuity property of $V^\star$, and we shall require the following Lipschitz assumptions on the original state transition and reward functions F and f.

($\mathbf{H}'_{\mathbf{lip}}$) There exists $K > 0$, such that for all $a \in A$, $e^0 \in E^0$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$, such that $\mathrm{pr}_2 \star \nu = \mathrm{pr}_2 \star \nu'$,

$$\begin{aligned}\mathbb{E}\big[d\big(F(x,a,\nu,\varepsilon^1_1,e^0), F(x',a,\nu',\varepsilon^1_1,e^0)\big)\big] &\le K\big(d(x,x') + \mathcal{W}(\mathrm{pr}_1 \star \nu, \mathrm{pr}_1 \star \nu')\big)\\ d(f(x,a,\nu), f(x',a,\nu')) &\le K\big(d(x,x') + \mathcal{W}(\mathrm{pr}_1 \star \nu, \mathrm{pr}_1 \star \nu)\big).\end{aligned}$$

Remark 4.5 Under Assumption ($\mathbf{H}'_{\mathbf{lip}}$), we see that once $\mathrm{pr}_i \star \nu = \mathrm{pr}_i \star \nu'$, $i = 1, 2$, then $F(x,a,\nu,\varepsilon^1_1,e^0) = F(x,a,\nu',\varepsilon^1_1,e^0)$, and $f(x,a,\nu) = f(x,a,\nu')$. In other words, the functions F and f depend on ν only through its marginal distribution on $\mathcal{X}$ and on A, and we shall write by misuse of notation: $F(x,a,\mu,\upsilon,e,e^0) = F(x,a,\mu \otimes \upsilon, e, e^0)$, $f(x,a,\mu,\upsilon) = f(x,a,\mu \otimes \upsilon)$, for $\mu \in \mathcal{P}(\mathcal{X})$, $\upsilon \in \mathcal{P}(A)$. Assumption ($\mathbf{H}'_{\mathbf{lip}}$) is then written as: there exists $K > 0$, such that for all $a \in A$, $e^0 \in E^0$, $x, x' \in \mathcal{X}$, $\mu, \mu' \in \mathcal{P}(\mathcal{X})$, $\upsilon \in \mathcal{P}(A)$,

$$\begin{aligned}\mathbb{E}\big[d\big(F(x,a,\mu,\upsilon,\varepsilon^1_1,e^0), F(x',a,\mu',\upsilon,\varepsilon^1_1,e^0)\big)\big] &\le K\big(d(x,x') + \mathcal{W}(\mu,\mu')\big)\\ d(f(x,a,\mu,\upsilon), f(x',a,\mu',\upsilon)) &\le K\big(d(x,x') + \mathcal{W}(\mu,\mu')\big).\end{aligned}$$

□

Proposition 4.3 *Assume that* ($\mathbf{H}'_{\mathbf{lip}}$) *holds true. Then, for all* $\gamma \in \big[0, \min\big(1, \frac{|\ln(\beta)|}{(\ln 2K)_+} - \eta\big)\big]$, *with* $\eta > 0$, $V^\star$ *is* γ*-Hölder with constant* $K_\gamma = \frac{2K\Delta_{\mathcal{X}}^{1-\gamma}}{1-\beta(2K)^\gamma}$:

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

Proof. Notice that $V^\star$ is the limiting point in $L^\infty(\mathcal{P}(\mathcal{X}))$ of the iterative sequence $V_{n+1} = \mathcal{T}V_n$, and we shall prove the γ-Hölder property by induction. Fix $\gamma \in [0,1]$. We start from $V_0 \equiv 0$ which is obviously γ-Hölder, and assume that at iteration n, V_n is γ-Hölder with a constant K_n. Under ($\mathbf{H}'_{\mathbf{lip}}$), (see Remark 4.5), recall the expression (4.10) of the Bellman operator $\mathcal{T}$:

$$\mathcal{T}W(\mu) = \sup_{\alpha_0 \in L^0(\Omega;A)} \mathbb{E}\Big[f\big(\xi, \alpha_0, \mu, \mathcal{L}(\alpha_0)\big) + \beta W\Big(\mathbb{P}^0_{F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon^0_1)}\Big)\Big], \quad \mu \in \mathcal{P}(\mathcal{X}),$$

for any $\xi \in L^0(\Omega;\mathcal{X})$ s.t. $\mathcal{L}(\xi)=\mu$. For any $\mu,\mu' \in \mathcal{P}(\mathcal{X})$, let us consider an optimal coupling $(\xi,\xi') \in L^0(\Omega;\mathcal{X})$ for the Wasserstein distance, i.e., $\mathbb{E}[d(\xi,\xi)] = \mathcal{W}(\mu,\mu')$. Under $(\mathbf{H}'_{\mathbf{lip}})$, we then have

$$\begin{aligned}|V_{n+1}(\mu) - V_{n+1}(\mu')| &\le 2K\mathcal{W}(\mu,\mu') \\ &\quad + \beta \sup_{\alpha_0 \in L^0(\Omega,;A)} \mathbb{E}\Big[\big|V_n\big(\mathbb{P}^0_{F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big) - V_n\big(\mathbb{P}^0_{F(\xi',\alpha_0,\mu',\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big)\big|\Big].\end{aligned}$$

Now, by induction hypothesis, we have

$$\begin{aligned}&\big|V_n\big(\mathbb{P}^0_{F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big) - V_n\big(\mathbb{P}^0_{F(\xi',\alpha_0,\mu',\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big)\big| \\ \le\ & K_n\big|\mathcal{W}\big(\mathbb{P}^0_{F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}, \mathbb{P}^0_{F(\xi',\alpha_0,\mu',\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big)\big|^\gamma,\end{aligned}$$

and by definition of the Wasserstein distance

$$\begin{aligned}\mathcal{W}\big(\mathbb{P}^0_{F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}, \mathbb{P}^0_{F(\xi',\alpha_0,\mu',\mathcal{L}(\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big) &\le \mathbb{E}\Big[d\big(F(\xi,\alpha_0,\mu,\mathcal{L}(\alpha_0),\varepsilon_1,e), F(\xi',\alpha_0,\mu',\mathcal{L}(\alpha_0),\varepsilon_1,e)\big)\Big]_{e:=\varepsilon_1^0} \\ &\le 2K\ \mathcal{W}(\mu,\mu'),\end{aligned}$$

where we used again $(\mathbf{H}'_{\mathbf{lip}})$ and the fact that $\mathbb{E}[d(\xi,\xi)] = \mathcal{W}(\mu,\mu')$. We then deduce that

$$\begin{aligned}\big|V_{n+1}(\mu) - V_{n+1}(\mu')\big| &\le 2K\mathcal{W}(\mu,\mu') + \beta K_n (2K)^\gamma \mathcal{W}(\mu,\mu')^\gamma \\ &\le (2K\Delta_{\mathcal{X}}^{1-\gamma} + \beta K_n (2K)^\gamma)\mathcal{W}(\mu,\mu')^\gamma,\end{aligned}$$

since $\mathcal{W}(\mu,\mu') \le \Delta_{\mathcal{X}}$ and $\gamma \le 1$. We then see that induction hypothesis at iteration $n+1$ holds true by setting $K_{n+1} := 2K\Delta_{\mathcal{X}}^{1-\gamma} + \beta K_n (2K)^\gamma$, which leads to the expression $K_n = 2K\Delta_{\mathcal{X}}^{1-\gamma}\frac{1-(\beta(2K)^\gamma)^n}{1-(\beta(2K)^\gamma)}$. Therefore, by taking $\gamma \in \big[0, \min\big(1, \frac{|\ln(\beta)|}{(\ln 2K)_+} - \eta\big)\big]$, with $\eta > 0$, we have $\beta(2K)^\gamma < 1$, and the sequence K_n converges to $K_\gamma := \frac{2K\Delta_{\mathcal{X}}^{1-\gamma}}{1-(\beta(2K)^\gamma)}$, which shows the required γ-Hölder property for $V^\star = \lim_n V_n$ with the constant K_γ. □

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

Lemma 4.6 (Quantization of $\mathcal{P}(\mathcal{X})$)
Fix $\eta > 0$. There exists a finite subset $\mathcal{M}_\eta = \{\mu^1,\dots,\mu^{N_\eta}\} \subset \mathcal{P}(\mathcal{X})$, such that for all $\mu \in \mathcal{P}(\mathcal{X})$, there exists $\mu^i \in \mathcal{M}_\eta$ satisfying $\mathcal{W}(\mu,\mu^i) \le \eta$.

Proof. As $\mathcal{X}$ is compact, there exists a finite subset $\mathcal{X}_\eta \subset \mathcal{X}$ such that $d(x,x_\eta) \le \eta/2$ for all $x \in \mathcal{X}$, where x_η denotes the projection of x on $\mathcal{X}_\eta$. Given $\mu \in \mathcal{P}(\mathcal{X})$, and $\xi \sim \mu$, we denote by ξ_η the quantization, i.e., the projection of ξ on $\mathcal{X}_\eta$, and by μ_η the discrete law of ξ_η. Thus, $\mathbb{E}[d(\xi,\xi_\eta)] \le \eta/2$, and therefore $\mathcal{W}(\mu,\mu_\eta) \le \eta/2$. The probability measure μ_η 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_η, in order to obtain another discrete probability measure μ_{η,n_η}. From the dual Kantorovich representation of Wasserstein distance, it is easy to see that for n_η large enough, $\mathcal{W}(\mu_\eta,\mu_{\eta,n_\eta}) \le \eta/2$, and so $\mathcal{W}(\mu,\mu_{\eta,n_\eta}) \le \eta$. We conclude the proof by noting that μ_{η,n_η} belongs to the set $\mathcal{M}_\eta$ of probability measures on $\mathcal{X}_\eta$ with weights valued in the finite grid G_η, hence $\mathcal{M}_\eta$ is a finite set of $\mathcal{P}(\mathcal{X}_\eta)$, of cardinal $N_\eta = n_\eta^{\#\mathcal{X}_\eta - 1}$. □

We can conclude this paragraph by showing the existence of an ϵ-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* $(\mathbf{H}'_{\mathbf{lip}})$ *holds true. Then, for all* $\epsilon > 0$*, there exists a lifted randomized feedback policy* $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$*, for some* $\mathfrak{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$*, that is* ϵ*-optimal for* $V^\star$*. Consequently, under* $\mathbf{Rand}(\xi, \mathcal{G})$*, the randomized feedback stationary control* $\alpha^{\mathfrak{a}_\epsilon} \in \mathcal{A}$ *is* $\frac{\epsilon}{1-\beta}$*-optimal for* $V(\xi)$*, and we have* $V(\xi) = \tilde{V}(\mu) = V^\star(\mu)$*, 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 $\mathcal{M}_\eta = \{\mu^1, \ldots, \mu^{N_\eta}\} \subset \mathcal{P}(\mathcal{X})$ as in Lemma 4.6, and an associated partition C^i_η, $i = 1, \ldots, N_\eta$, of $\mathcal{P}(\mathcal{X})$, satisfying

$$C^i_\eta \subset\ B_\eta(\mu^i) := \Big\{\mu \in \mathcal{P}(\mathcal{X}) : \mathcal{W}(\mu, \mu^i) \leq \eta\Big\}, \quad i = 1, \ldots, N_\eta.$$

For any μ^i, $i = 1, \ldots, N_\eta$, and by (4.13), there exists $\mathrm{a}^i_\epsilon \in L^0(\mathcal{X} \times [0,1]; A)$ such that

$$V^\star(\mu^i) \leq [\mathbb{T}^{\mathrm{a}^i_\epsilon} V^\star](\mu^i) + \frac{\epsilon}{3}. \tag{4.14}$$

From the partition C^i_η, $i = 1, \ldots, N_\eta$ of $\mathcal{P}(\mathcal{X})$, associated to $\mathcal{M}_\eta$, we construct the function $\mathfrak{a}_\epsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to A$ as follows. Let 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

$$\mathfrak{a}_\epsilon(\mu, x, u) \quad = \quad \mathrm{a}^i_\epsilon\big(\zeta(\mu, \mu^i, x, h_1(u)), h_2(u)\big), \quad \text{when } \mu \in C^i_\eta,\ i = 1, \ldots, N_\eta,\ x \in \mathcal{X}, u \in [0,1],$$

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

$$[\mathcal{T}^{\boldsymbol{\pi}_\varepsilon} V^\star](\mu^i) = \ [\mathbb{T}^{\mathrm{a}^i_\epsilon} V^\star](\mu^i), \quad i = 1, \ldots, N_\eta, \tag{4.15}$$

by (4.12). Let us now check that such $\boldsymbol{\pi}_\epsilon$ yields an ϵ-optimal randomized feedback policy for η 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 $\mu \in C^i_\eta$, $i = 1, \ldots, N_\eta$, and $\xi_\eta := \zeta(\mu, \mu_\eta, \xi, U_1)$. Observe by Lemma 4.6 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 $\mu \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - V^\star(\mu) = \ & \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta)\Big) + \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta) - V^\star(\mu_\eta)\Big) \\
& + \big(V^\star(\mu_\eta) - V^\star(\mu)\big) \\
\geq \ & \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta)\Big) - \frac{\epsilon}{3} - \frac{\epsilon}{3},
\end{aligned} \tag{4.16}$$

where we used (4.14)-(4.15) and the fact that $|V^\star(\mu_\eta) - V^\star(\mu)| \leq \epsilon/3$ for η small enough by uniform continuity of $V^\star$ in Proposition 4.3. Moreover, by observing that $\mathfrak{a}_\epsilon(\mu, \xi, U) = \mathfrak{a}_\epsilon(\mu_\eta, \xi_\eta, U_2) =: \alpha_0$, so that $\boldsymbol{\pi}_\epsilon(\mu) = \mathcal{L}(\xi, \alpha_0)$, $\boldsymbol{\pi}_\epsilon(\mu_\eta) = \mathcal{L}(\xi_\eta, \alpha_0)$, we have

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) = \ & \mathbb{E}\Big[f(Y) + \beta V^\star(\mathbb{P}^0_{F(Y, \varepsilon_1, \varepsilon^0_1)})\Big], \\
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta) = \ & \mathbb{E}\Big[f(Y_\eta) + \beta V^\star(\mathbb{P}^0_{F(Y_\eta, \varepsilon_1, \varepsilon^0_1)})\Big],
\end{aligned}$$

where $Y = (\xi, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu))$, and $Y_\eta = (\xi_\eta, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu_\eta))$. Under $(\mathbf{H}'_{\mathbf{lip}})$, by using the γ-Hölder property of $V^\star$ with constant K_γ in Proposition 4.3, and by definition of the Wasserstein distance (recall that $\xi \sim \mu$, $\xi_\eta \sim \mu_\eta$), we then get

$$\begin{aligned}\big|[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu_\eta)\big| &\leq 2K\mathbb{E}\big[d(\xi,\xi_\eta)\big] \\ &\quad +\beta K_\gamma \mathbb{E}\Big[\mathbb{E}\big[d\big(F(\xi,\alpha_0,\boldsymbol{\pi}_\epsilon(\mu),\varepsilon_1,e),F(\xi_\eta,\alpha_0,\boldsymbol{\pi}_\epsilon(\mu_\eta),\varepsilon_1,e)\big)\big]_{e:=\varepsilon_1^0}\Big] \\ &\leq 2K(1+\beta K_\gamma)\mathbb{E}\big[d(\xi,\xi_\eta)\big].\end{aligned}$$

Now, by the coupling Lemma 4.1, $\mathbb{E}\big[d(\xi,\xi_\eta)\big]$ can be made as small as possible provided that $\eta \geq \mathcal{W}(\mu,\mu_\eta)$ is small enough, and therefore, $\big|[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu_\eta)\big| \leq \epsilon/3$. Plugging into (4.16), we obtain $\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star(\mu) - V^\star(\mu) \geq -\epsilon$, for all $\mu \in \mathcal{P}(\mathcal{X})$, which means that $\boldsymbol{\pi}_\epsilon$ is ϵ-optimal for $V^\star$. The rest of the assertions in the Theorem follow from the verification result in Proposition 4.2. □

Remark 4.6 We stress the importance of the coupling Lemma in the construction of ϵ-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 $[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu)$ and $[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu_\eta)$ close to each other is to couple terms to have the same control in F and f. This is achieved by turning μ into μ_η, ξ into ξ_η and set $\alpha_0 = \mathfrak{a}_\epsilon(\mu,\xi,U) = \mathfrak{a}_\epsilon(\mu_\eta,\xi_\eta,U_2)$. Turning μ into μ_η is a simple quantization, but turning ξ into ξ_η is obtained thanks to the coupling Lemma. □

4.4 Relaxing the randomization hypothesis

We consider the general case by relaxing the randomization hypothesis and by assuming only 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 4.7 *Assume that* $(\mathbf{H}'_{\mathbf{lip}})$ *holds true. Then, for all* $\gamma \in \big[0, \min\big(1, \frac{|\ln(\beta)|}{(\ln 2K)_+} - \eta\big)\big]$, *with* $\eta > 0$, *and setting* $K_\gamma = \frac{2K\Delta_{\mathcal{X}}^{1-\gamma}}{1-\beta(2K)^\gamma}$, *we have*

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

Consequently, V is γ-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. $X_t^{\xi,\alpha} = f_{t,\pi}(\xi,\Gamma,(\varepsilon_s)_{s\leq t},(\varepsilon_s^0)_{s\leq t})$, thus $\mathbb{P}^0_{X_t^{\xi,\alpha}} = \mathcal{L}(f_{t,\pi}(\xi,\Gamma,(\varepsilon_s)_{s\leq t},(e_s^0)_{s\leq t}))_{e_s^0=\varepsilon_s^0,s\leq t}$. We thus have

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

and so

$$\mathbb{E}\Big[\mathcal{W}(\mathbb{P}^0_{X_t^{\xi,\alpha}},\mathbb{P}^0_{X_t^{\xi',\alpha}})\Big] \leq \mathbb{E}\big[d(X_t^{\xi,\alpha},X_t^{\xi',\alpha})\big], \tag{4.17}$$

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

$$|V^\alpha(\xi) - V^\alpha(\xi')| \leq 2K \sum_{t=0}^{\infty} \beta^t \mathbb{E}\big[d(X_t^{\xi,\alpha}, X_t^{\xi',\alpha})\big]. \tag{4.18}$$

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

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

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

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

where the last inequality comes from (4.17). Denoting by $\delta_t(\xi,\xi') := \sup_{\alpha\in\mathcal{A}} \mathbb{E}\big[d(X_t^{\xi,\alpha}, X_t^{\xi',\alpha})\big]$, and noting that $\delta_0(\xi,\xi') = |\xi - \xi'|$, it follows by induction that

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

By the same arguments as in Theorem 2.1, and choosing γ as in the assertion of the theorem, we obtain that

$$\sum_{t=0}^{\infty} \beta^t \delta_t(\xi,\xi') \leq \frac{\Delta_{\mathcal{X}}^{1-\gamma}}{1-\beta(2K)^\gamma} \big(\mathbb{E}\big[d(\xi,\xi')\big]\big)^\gamma, \quad \xi,\xi' \in L^0(\mathcal{G};\mathcal{X}),$$

and conclude with (4.18). □

Theorem 4.2 *Assume that $\mathcal{G}$ is rich enough and $(\mathbf{H}'_{\mathbf{lip}})$ 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 ϵ-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_η 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 $\mathbf{Rand}(\xi',\mathcal{G}')$ holds true, and we then have $V(\xi') = \tilde{V}(\mu)$ from Theorem 4.1. Consider now the quantized random variables ξ_η and ξ'_η, which have the same law, and satisfy respectively the randomization hypothesis $\mathbf{Rand}(\xi_\eta,\mathcal{G})$ and $\mathbf{Rand}(\xi'_\eta,\mathcal{G}')$ 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 η to zero, that $V(\xi) = V(\xi')$, and thus $V(\xi) = \tilde{V}(\mu)$, which proves the required result.

Finally, the existence of an ϵ-optimal control for $V(\xi)$ is obtained as follows. From the uniform continuity of V in Lemma 4.7, there exists η 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 ϵ-optimal (randomized feedback stationary) control for $V(\xi)$. □

Remark 4.7 From Theorems 4.1 and 4.2, under the condition that $\mathcal{G}$ is rich enough and $(\mathbf{H}'_{\mathbf{lip}})$ 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 $\mathcal{T}^{\mathfrak{a}} := \mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}}$ equal to

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

with $Y^{\mathfrak{a}}(\mu, x, u) := (x, \mathfrak{a}(\mu, x, u), \boldsymbol{\pi}^{\mathfrak{a}}(\mu))$, and $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$. □

Remark 4.8 ($\mathcal{G}$ rich enough and dynamic programming) The condition that $\mathcal{G}$ is rich enough is crucial for obtaining the Bellman fixed point equation for the value function V. Let us illustrate this fact with a counter-example similar to Example 3.1 in [11]. Consider $\mathcal{X} = \{-1, 1\} = A$, $\varepsilon_1 \sim \mathcal{B}(1/2)$, $F(x, a, \nu, e, e^0) = ax$, $f(x, a, \nu) = -\mathcal{W}(\mathrm{pr}_1 \star \nu, \mathcal{B}(1/2))$. 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 (δ_{-1} or δ_1). Assume that $\Gamma = 1$ a.s., so that that $\mathcal{G}$ is the trivial σ-algebra. In this case, $\xi =: x$ and α_0 are then necessarily determinist, and thus $X_1^{\xi,\alpha} = \alpha_0 \xi$ has to be determinist 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 $\mathbb{P}^0_{X_t^{x,\alpha}} = \delta_x$ for $t = 0, 1$, and $\mathbb{P}^0_{X_t^{x,\alpha}} = \mathcal{B}(1/2)$ afterwards. This control is clearly optimal, and the associated gain is $V(\xi) = -\frac{1+\beta}{2}$. If V satisfied the DPP, we would have $V(x) = \sup_{a \in A}(-1/2 + \beta V(ax))$, which is equivalent to $-\frac{1+\beta}{2} = -\frac{1}{2} - \beta\frac{1+\beta}{2}$, and this is clearly false.

Intuitively, the reason why the DPP is not satisfied in the previous example is that from time $t = 1$ we have the possibility to randomize actions using ε_1, whereas at time $t = 0$ no randomization was possible (even by quantizing, because $\mathcal{G}$ is not rich enough). As discussed in the next remark, the possibility to randomize implies that the game from time 1 is more advantageous than the game from time 0, which contradicts the idea behind the DPP that from time 1, we are in the same game, but only starting from a different state. □

Remark 4.9 (Open-loop vs feedback controls and randomization hypothesis) In Theorem 4.1, we build under randomization hypothesis, an ϵ-optimal control of the form $\alpha_t = \mathfrak{a}_\epsilon(\mathbb{P}^0_{X_t}, X_t, U_t)$ (randomized feedback controls), for any $\epsilon > 0$, which implies that optimizing over open-loop controls gives the same optimal value than optimizing over randomized feedback controls. In standard (non-McKean-Vlasov) MDPs, it is well known that we can even restrict to controls of the form $\alpha_t = \mathfrak{a}(X_t)$, i.e., (not randomized) feedback controls. Randomizing actions is thus not necessary in standard MDPs. However, in the case of McKean-Vlasov MDPs, it can be crucial to optimize our gain (even when the reward does not depend upon the law of the control). To illustrate this, let us consider the same example as in Remark 4.8, and assume that we start from $\xi = 1$ a.s.

- If we use a feedback control (not randomized), it is clear that the law of X_t will always be a Dirac, and thus the gain will be equal to $\sum_{t=0}^{\infty} \beta^t(-\frac{1}{2}) = -\frac{1}{2(1-\beta)}$ which is the worst possible gain.

- However, with open-loop controls, even when the randomization hypothesis $\mathbf{Rand}(\xi, \mathcal{G})$ is not satisfied, if at some point we can use the past noise to randomize (even just slightly), there will be some times where the law of the state is not a Dirac and the gain will then be strictly greater than $-\frac{1}{2(1-\beta)}$. In our case, we have seen in the previous remark that we could reach $V = -\frac{1+\beta}{2}$.

- Finally, if $\mathcal{G}$ was rich enough, for instance if $\Gamma \sim \mathcal{U}([0,1]))$, then $\mathbf{Rand}(\xi, \mathcal{G})$ would hold true, and we could freely randomize from the beginning by choosing $\alpha_0 = \mathbf{1}_{G<1/2} \sim \mathcal{B}(1/2)$ (thus $X_1 \sim \mathcal{B}(1/2)$), and then, fixing $\alpha_t = 1$ for all $t \in \mathbb{N}_\star$, we would have $X_t \sim \mathcal{B}(1/2)$ for all $t \in \mathbb{N}^\star$. With this strategy, the total gain would be equal to $-\frac{1}{2}$, which is the best possible gain given that we start from $\xi = 1$ a.s.

To summarize, this example shows that the optimal value under $\mathbf{Rand}(\xi, \mathcal{G})$ is here strictly greater than the optimal value over open-loop controls without $\mathbf{Rand}(\xi, \mathcal{G})$, which is itself strictly greater than the optimal value over feedback controls. This highlights that in general the supremum over "randomized feedback" $\geq$ "open-loop" $\geq$ "feedback" controls, and that the inequalities can be strict, the underlying idea being that the more (and sooner) we can randomize, the better it is. We finally point out that the randomization hypothesis $\mathbf{Rand}(\xi, \mathcal{G})$ is natural, since in practice, nothing prevents us from using (pseudo-)uniform variables in our strategies to randomize our actions. □

4.5 Computing value function and ϵ-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 design two methods for computing the value function and optimal strategies:

(a) Value iteration. We approximate the value function $V = \tilde{V} = V^\star$ 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 $\mathfrak{a}_N$ by (recall Remark 4.7)

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

From $\mathfrak{a}_N$, we then construct an approximate randomized feedback stationary control $\alpha^{\mathfrak{a}_N}$ according to the procedure described in Remark 4.3.

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

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

- Greedy strategy: we compute

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

We stop at iteration K to obtain $\mathfrak{a}_K$, and then construct an approximate randomized feedback control $\alpha^{\mathfrak{a}_K}$ according to the procedure described in Remark 4.3.

Practical computation. Since a randomized feedback control α is a measurable function $\mathfrak{a}$ of $(\mathbb{P}^0_{X_t^{\xi,\alpha}}, X_t^{\xi,\alpha}, U_t)$, 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 $\mathcal{X}_\eta = \{x^1, \ldots, x^{N_\eta}\} \subset \mathcal{X}$ associated with a partition B^i_η, $i = 1, \ldots, N_\eta$, of $\mathcal{X}$, satisfying: $B^i_\eta \subset \big\{x \in \mathcal{X} : d(x, x^i) \leq \eta\big\}$, $i = 1, \ldots, N_\eta$, with $\eta > 0$. For any $x \in \mathcal{X}$, we denote by $[x]_\eta$ (or simply x_η) its projection on $\mathcal{X}_\eta$, defined by: $x_\eta = x^i$, for $x \in B^i_\eta$, $i = 1, \ldots, N_\eta$.

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*

$$X^{\eta,\xi,\alpha}_{t+1} = \big[F(X^{\eta,\xi,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\eta,\xi,\alpha}_t, \alpha_t)}, \varepsilon_{t+1}, \varepsilon^0_{t+1})\big]_\eta, \quad t \in \mathbb{N}, \;\; X^{\eta,\xi,\alpha}_0 = \xi,$$

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

$$V^\alpha_\eta(\xi) = \mathbb{E}\Big[\sum_{t=0}^\infty \beta^t f\big(X^{\eta,\xi,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\eta,\xi,\alpha}_t, \alpha_t)}\big)\Big].$$

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 ϵ-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 $(\mathbf{H}'_{\mathbf{lip}})$ holds true. Fix $\xi \in L^0(\mathcal{G}; \mathcal{X})$. Given $\eta > 0$, let us define ξ_η the projection of ξ 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 ξ_η. For $\epsilon > 0$, let $\mathfrak{a}_\epsilon$ be a randomized feedback policy that is ϵ-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 $\alpha^{\eta,\epsilon}_t = \mathfrak{a}_\epsilon(\mathbb{P}^0_{X^{\eta,\epsilon}_t}, X^{\eta,\epsilon}_t, U_{\eta,t})$, $t \in \mathbb{N}$, where we set $X^{\eta,\epsilon}_t := X^{\eta,\xi_\eta,\alpha^{\epsilon,\eta}}_t$. Then, for any $\delta > 0$ and $\gamma \in \big[0, \min\big(1, \frac{|\ln\beta|}{(\ln 2K)_+} - \delta\big)\big]$, the control $\alpha^{\eta,\epsilon}$ is $\mathcal{O}(\eta^\gamma + \epsilon)$-optimal for the CMKV-MDP X with initial state ξ.*

Proof. *Step 1.* For $\delta > 0$ and $\gamma \in \big[0, \min\big(1, \frac{|\ln\beta|}{(\ln 2K)_+} - \delta\big)\big]$, let us show that

$$\sup_{\alpha\in\mathcal{A}} \sum_{t=0}^\infty \beta^t \mathbb{E}\big[d(X^{\xi,\alpha}_t, X^{\eta,\xi_\eta,\alpha}_t)\big] \leq \; C\eta^\gamma, \tag{4.19}$$

for some constant C that depends only on K, β and γ. 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 $t \in \mathbb{N}$,

$$\mathbb{E}\big[d(X^{\xi,\alpha}_{t+1}, X^{\eta,\xi_\eta,\alpha}_{t+1})\big] \leq \; \eta + \mathbb{E}\big[\Delta\big(X^{\xi,\alpha}_t, X^{\eta,\xi_\eta,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\xi,\alpha}_t,\alpha_t)}, \mathbb{P}^0_{(X^{\eta,\xi_\eta,\alpha}_t,\alpha_t)}, \varepsilon^0_{t+1}\big)\big],$$

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

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

by the same argument as in (4.17). Hence, the sequence $(\mathbb{E}\big[d(X_t^{\xi,\alpha},X_t^{\eta,\xi_\eta,\alpha})\big])_{t\in\mathbb{N}}$ satisfies the same type of induction inequality as in (2.16) in Theorem 2.1 with η instead of M_N, and thus the same derivation leads to the required result (4.19). From the Lipschitz condition on f, we deduce by the same arguments as in (4.17) in Lemma 4.7 that

$$\sup_{\alpha\in\mathcal{A}}\big|V^\alpha(\xi_\eta) - V_\eta^\alpha(\xi_\eta)\big| = \mathcal{O}(\eta^\gamma). \tag{4.20}$$

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

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

The first and last terms I_1 and I_4 are smaller than $\mathcal{O}(\eta^\gamma)$ by the γ-Hölder property of V^α and V in Lemma 4.7. By (4.20), the second term I_2 is of order $\mathcal{O}(\eta^\gamma)$ as well for η 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 $\mathfrak{a}_\epsilon$ for the discretized CMKV-MDP. Denote by $\boldsymbol{\pi}_\epsilon$ the lifted randomized feedback policy associated to $\mathfrak{a}_\epsilon$, and recall by Remark 4.3 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

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

$Y^{\mathfrak{a}}(\mu,x,u) = (x,\mathfrak{a}(\mu,x,u),\boldsymbol{\pi}^{\mathfrak{a}}(\mu))$ 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

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

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

$$\Big|[\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta) - [\mathcal{T}^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta)\Big| \le \beta\mathbb{E}\Big[\big|\tilde{V}\big(\mathbb{P}^0_{[F(Y_\eta,\varepsilon_1,\varepsilon_1^0)]_\eta}\big) - \tilde{V}\big(\mathbb{P}^0_{F(Y_\eta,\varepsilon_1,\varepsilon_1^0)}\big)\big|\Big],$$

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

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

From the γ-Hölder property of $\tilde{V}$ in Proposition 4.3, we deduce that $I_3^2 = \mathcal{O}(\eta^\gamma)$. Finally, for I_3^1, since $\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}$ is a β-contracting operator on $(L^\infty(\mathcal{M}_\eta), \|\cdot\|_{\eta,\infty})$, we have

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

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. □

Remark 4.10 Back to the N-agent MDP problem with open-loop controls, recall from Section 2, that it suffices to find an ϵ-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 ϵ-optimal control α^ϵ given by Proposition 4.4 can be associated to an ϵ-optimal open-loop policy π^ϵ such that $\alpha^\epsilon = \alpha^{\pi^\epsilon}$ (where π_t^ϵ 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 $\alpha_t^{\epsilon,i}$ 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 ξ by ξ^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 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

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

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:

$$Q(\mu, \hat{a}) = \hat{f}(\mu, \hat{a}) + \beta\mathbb{E}\Big[\sup_{\hat{a}'\in\hat{A}(\mathcal{X})} Q\big(\mu_1^{\hat{a}}, \hat{a}'\big)\Big], \tag{4.21}$$

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 [11] (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 (4.21) 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 A.3, one can associate to such probability kernels $\hat{a}_n^\mu$, a function $\mathfrak{a}_n : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to A$, such that $\mathcal{L}(\mathfrak{a}_n(\mu, x, U)) = \hat{a}_n^\mu(x)$, $\mu \in \mathcal{P}(\mathcal{X})$, $x \in \mathcal{X}$, 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 4.6 in order to reduce the learning problem to a finite-dimensional problem for the computation of an approximate optimal randomized feedback policy $\mathfrak{a}_n$ for the CMKV-MDP. □

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 dynamic programming 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 $(S, \mathcal{S})$, $(T, \mathcal{T})$ be two measurable spaces, then*

(1) the map $x \in (S, \mathcal{S}) \mapsto \delta_x \in (\mathcal{P}(S), \mathcal{C}(S))$ is measurable.

(2) The map $(\mu, \nu) \in (\mathcal{P}(S), \mathcal{C}(S)) \times (\mathcal{P}(T), \mathcal{C}(T)) \mapsto \mu \otimes \nu \in (\mathcal{P}(S \times T), \mathcal{C}(S \times T))$ is measurable.

(3) If $F \in L^0(S; T)$, then the map $\mu \in (\mathcal{P}(S), \mathcal{C}(S)) \mapsto F \star \mu \in (\mathcal{P}(T), \mathcal{C}(T))$ is measurable.

Proof. (1). Fix $A \in \mathcal{S}$, then $x \mapsto \delta_x(A) = \mathbf{1}_A(x)$ is clearly a measurable function. (2). Fix $(A, B) \in \mathcal{S} \times \mathcal{T}$, then $(\mu, \nu) \mapsto (\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ is a measurable function. (3). Fix $B \in \mathcal{T}$, then $\mu \mapsto (F \star \mu)(B) = \mu(F^{-1}(B))$ is a measurable function. □

Corollary 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 $\hat{F} : (\mathcal{P}(S), \mathcal{C}(S)) \times (T, \mathcal{T}) \to (\mathcal{P}(U), \mathcal{C}(U))$ given by $\hat{F}(\mu, x) := F(\cdot, x) \star \mu$ is measurable.*

Proof. This is a consequence of Lemma A.1 and the measurable composition $(\mu, x) \mapsto (\mu, \delta_x) \mapsto \mu \otimes \delta_x \mapsto F \star (\mu \otimes \delta_x) = \hat{F}(\mu, x)$. □

Definition A.1 (Conditional law) *Fix two measurable spaces $(S,\mathcal{S})$ and $(T,\mathcal{T})$, and let X,Y be two random variables on $(\Omega,\mathcal{F},\mathbb{P})$ valued respectively on S and T. Then a conditional law of Y knowing X is a $\sigma(X)$-measurable $(\mathcal{P}(T),\mathcal{C}(T))$-valued random variable $\mathbb{P}_Y^X$, also denoted $\mathcal{L}(Y|X)$, such that*

$$\mathbb{P}_Y^X(A) = \mathbb{P}[Y \in A \mid X], \quad \forall A \in \mathcal{T} \quad a.s.$$

Furthermore, given $x \in S$, the conditional law of Y knowing $X = x$, denoted $\mathbb{P}_Y^{X=x}$ or $\mathcal{L}(Y|X=x)$, is the image of x by any probability kernel $\hat{a} \in \hat{T}(S)$ such that $\mathcal{L}(Y|X) = \hat{a}(X)$ a.s.

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 $Z \perp\!\!\!\perp X$ is a random variable valued in a measurable space V and $\varphi : S \times V \to T$ 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 $(\mathcal{P}(T),\mathcal{C}(T))$.*

Proof. The first assertion is stated in Theorem 5.3 in [12]., and the second one follows from Fubini's theorem. □

Corollary A.2 *Fix a Borel space $(S,\mathcal{S})$ and a measurable space $(T,\mathcal{T})$. Then for any joint law $\boldsymbol{\pi} \in \mathcal{P}(S \times T)$ with marginal law $\mu \in \mathcal{P}(S)$, there exists a probability kernel ν from S to $\mathcal{P}(T)$ s.t. $\boldsymbol{\pi} = \mu \cdot \nu$.*

Lemma A.3 (Kernels and randomization) *Fix a Borel space $(S,\mathcal{S})$ and a measurable space $(T,\mathcal{T})$. For any probability kernel ν from S to $\mathcal{P}(T)$, there exists a measurable function $\phi : S \times [0,1] \to T$ s.t. $\nu(s) = \mathcal{L}(\phi(s,U))$, for all $s \in S$, where U is a uniform random variable.*

Proof. See Lemma 2.22 in [12]. □

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 $\sigma(\xi,\Gamma,(\varepsilon)_{s\le t},(\varepsilon_s^0)_{s\le t})$-measurable, and $\mathbb{P}^0_{(X_t^{\xi,\alpha},\alpha_t)}$ 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

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

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

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 $f : E \times F \to G$. Then*

$$\mathcal{W}\Big(\frac{1}{N}\sum_{i=1}^{N}\delta_{f(X,U_i)}, \mathbb{P}^X_{f(X,U_1)}\Big) \quad \underset{N\to\infty}{\overset{a.s.}{\to}} \quad 0.$$

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. □

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*

$$\Big(F^{-1}(U), F(F^{-1}(U)) - U\Big) \overset{d}{=} (F^{-1}(U), V\Delta F(F^{-1}(U)),$$

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

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

The second term can be decomposed as

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

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

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

which proves the result. □

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. *ϕ and ϕ^{-1} are uniformly continuous,*
2. *for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have $Im\left(F^{-1}_{\phi\star\mu}\right) \subset Im(\phi)$. In particular, $\phi^{-1} \circ F^{-1}_{\phi\star\mu}$ 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 $\mathrm{d}((u_n)_{n\in\mathbb{N}}, (v_n)_{n\in\mathbb{N}}) := \sum_{n\geq 0} \frac{1}{2^n}|u_n - v_n|$. ϕ_1 is clearly injective and uniformly continuous (even Lipschitz). The compactness of $\mathcal{X}$ implies that its inverse ϕ_1^{-1} is uniformly continuous as well. Let us now consider $\phi_2 : ([0,1]^{\mathbb{N}}, \mathrm{d}) \mapsto [0,1]$ 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 $\iota : \mathbb{N} \to \mathbb{N}^2$ 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...). ϕ_2 is clearly injective, uniformly continuous, as well as its inverse ϕ_2^{-1}. Thus, $\phi := \phi_2 \circ \phi_1$ defines an embedding of $\mathcal{X}$ into $\mathbb{R}$, such that ϕ and ϕ^{-1} are uniformly continuous.
2. $F^{-1}_{\phi\star\mu}$ being caglad, and $\mathrm{Im}(\phi)$ being closed (by compactness of $\mathcal{X}$), it is enough to prove that $F^{-1}_{\phi\star\mu}(u) \in \mathrm{Im}\phi$ for almost every $u \in [0,1]$ (in the Lebesgue sense). However, given a uniform variable U, we have $F^{-1}_{\phi\star\mu}(U) \sim \phi \star \mu$, and thus

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

□

Proof of Lemma 4.1
(1) We first consider the case where $\mathcal{X} \subset \mathbb{R}$. Let us call F_μ the distribution function of $\mu \in \mathcal{P}(\mathcal{X})$, and F_μ^{-1} its generalized inverse. Let us define the function $\zeta : \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1] \to \mathcal{X}$ by

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

which is measurable by noting that the measurability in μ, μ' comes from the continuity of

$$\begin{array}{rcl} \mathcal{P}(\mathcal{X}) & \to & L^1_{caglad}([0,1], \mathcal{X}) \\ \mu & \mapsto & F_\mu^{-1}. \end{array}$$

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

$$\begin{array}{rcl} (\xi, \zeta(\mu, \mu', \xi, V)) & = & (\xi, F_{\mu'}^{-1}\big(F_\mu(\xi) - V\Delta F_\mu(\xi)\big)) \\ & \overset{d}{=} & (F_\mu^{-1}(U), F_{\mu'}^{-1}\big(F_\mu(F_\mu^{-1}(U)) - V\Delta F_\mu(F_\mu^{-1}(U))\big)) \\ & = & (F_\mu^{-1}(U), F_{\mu'}^{-1}\big(F_\mu(F_\mu^{-1}(U)) - V\Delta F_\mu(F_\mu^{-1}(U))\big)) \\ & \overset{d}{=} & \big(F_\mu^{-1}(U), F_{\mu'}^{-1}(U)\big), \end{array}$$

where the last equality holds by Lemma C.1. It is well-known (see e.g. Theorem 3.1.2 in [17]) that $\big(F_\mu^{-1}(U), F_{\mu'}^{-1}(U)\big)$ is an optimal coupling for (μ, μ'), and so $\mathcal{W}(\mu, \mu') = \mathbb{E}\big[d(\xi, \zeta(\mu, \mu', \xi, V))\big]$.

(2) Let us now consider the case of a general compact Polish space $\mathcal{X}$. Denoting by $\zeta_{\mathbb{R}}$ the "ζ" 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 ξ, and define $\xi' := \zeta(\mu, \mu', \xi, U)$. By definition of ζ, its clear that $\xi' \sim \mu'$, and

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

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

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

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

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

so that we can choose $\delta = \frac{\gamma}{\Delta_{\mathcal{X}}}\frac{\epsilon}{2}$. On the other hand, by uniform continuity of ϕ 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.2), we thus conclude that $d(\mu, \mu') < \eta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon$. □

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.* 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, D. Possamai, and X. Tan. McKean-Vlasov optimal control: the dynamic programming principle. arXiv:1907.08860,, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] H. Gu, X. Guo, X. Wei, and R. Xu. Dynamic programming principles for learning MFCs. arXiv:1911.07314, 2019.

[12] O. Kallenberg. *Foundations of Modern Probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

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

[14] M. Laurière and O. Pironneau. Dynamic programming for mean-field type controls. *Journal of Optimization Theory*, 169(3):902–924, 2016.

[15] H. Pham and X. Wei. Discrete time McKean-Vlasov control problem: a dynamic programming approach. *Applied Mathematics and Optimization*, 74(3):487–506, 2016.

[16] 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.

[17] S.T. Rachev and L. Rüschendorf. *Mass transportation problems*. Springer Verlag, 1998.

[18] R.S. Sutton and A.G. Barto. *Reinforcement learning: an introduction*. Cambridge, MA, 2017, 2nd edition.

[19] A. Van der Vaart and J.A. Wellner. *Weak convergence of empirical processes*. Springer Verlag, 1996.

[20] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.