Quantitative propagation of chaos for mean field Markov decision process with common noise

Médéric Motte, Huyên Pham

Quantitative propagation of chaos for mean field Markov decision process with common noise

Médéric MOTTE * Huyên PHAM †

July 25, 2022

Abstract

We investigate propagation of chaos for mean field Markov Decision Process with common noise (CMKV-MDP), and when the optimization is performed over randomized open-loop controls on infinite horizon. We first state a rate of convergence of order M_N^γ, where M_N is the mean rate of convergence in Wasserstein distance of the empirical measure, and $\gamma \in (0,1]$ is an explicit constant, in the limit of the value functions of N-agent control problem with asymmetric open-loop controls, towards the value function of CMKV-MDP. Furthermore, we show how to explicitly construct $(\epsilon + \mathcal{O}(M_N^\gamma))$-optimal policies for the N-agent model from ϵ-optimal policies for the CMKV-MDP. Our approach relies on sharp comparison between the Bellman operators in the N-agent problem and the CMKV-MDP, and fine coupling of empirical measures.

1 Introduction

We consider a social planner problem with N cooperative agents in a mean-field discrete time model with common noise over an infinite horizon. The controlled state process $\boldsymbol{X} = (X^i)_{i\in[\![1,N]\!]}$ of the N-agent model is given by the dynamical random system

$$\left\{ \begin{array}{rcl} X_0^i &=& x_0^i, \\ X_{t+1}^i &=& F(X_t^i, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(X_t^j,\alpha_t^j)}, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}. \end{array} \right. \tag{1.1}$$

Here, x_0^i, $i \in [\![1,N]\!]$, are the initial states valued in a compact Polish space $\mathcal{X}$ with metric d, $(\varepsilon_t^i)_{i\in[\![1,N]\!]t\in\mathbb{N}^\star}$ is a family of mutually i.i.d. random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, valued in some measurable space E, and representing idiosyncratic noises, while $(\varepsilon_t^0)_{t\in\mathbb{N}^\star}$ is another family of i.i.d. random variables valued in some measurable space E^0, and representing the common noise (independent of idiosyncratic noise). The control α^i followed by agent i, is a process, valued in some compact Polish space A with metric d_A, and adapted with respect to the filtration $(\mathcal{F}_t^N)_{t\in\mathbb{N}}$ generated by $\boldsymbol{\varepsilon} = \big((\varepsilon_t^i)_{i\in[\![1,N]\!]}, \varepsilon_t^0\big)_{t\in\mathbb{N}^\star}$ and also completed with a family of

*LPSM, Université Paris Cité medericmotte at gmail.com

†LPSM, Université Paris Cité, and CREST-ENSAE, pham at lpsm.paris The author acknowledges support of the ANR 18-IDEX-001. This work was also partially supported by the Chair Finance & Sustainable Development / the FiME Lab (Institut Europlace de Finance)

mutually i.i.d. uniform random variables $\boldsymbol{U} = (U_t^i)_{i\in[\![1,N]\!],t\in\mathbb{N}}$ that are used for randomization of the controls. The mean-field interaction between the agents is formalized via the state transition function F by the dependence upon the empirical measure of both state/action of all the other agents: here F is a measurable function from $\mathcal{X}\times A\times\mathcal{P}(\mathcal{X}\times A)\times E\times E^0$ into $\mathcal{X}$, where $\mathcal{P}(\mathcal{X}\times A)$ is the space of probability measures on the product space $\mathcal{X}\times A$.

The objective of the social planner is to maximize over the set $\boldsymbol{\mathcal{A}}$ of A^N-valued $(\mathcal{F}_t^N)_{t\in\mathbb{N}}$-adapted processes $\boldsymbol{\alpha} = (\alpha_t^i)_{i\in[\![1,N]\!],t\in\mathbb{N}}$ a criterion in the form

$$V_N^{\boldsymbol{\alpha}}(\boldsymbol{x}_0) \;:=\; \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N\sum_{t=0}^\infty \beta^t f\big(X_t^i,\alpha_t^i,\frac{1}{N}\sum_{j=1}^N\delta_{(X_t^j,\alpha_t^j)}\big)\Big],$$

where we set $\boldsymbol{x}_0 = (x_0^i)_{i\in[\![1,N]\!]}\in\mathcal{X}^N$ for the initial state of the N agent system. Here $\beta\in(0,1)$ is a discount factor, and f is a bounded measurable real-valued function on $\mathcal{X}\times A\times\mathcal{P}(\mathcal{X}\times A)$. The value function for this optimization problem is defined on $\mathcal{X}^N$ as

$$V_N(\boldsymbol{x}_0) := \sup_{\boldsymbol{\alpha}\in\boldsymbol{\mathcal{A}}} V_N^{\boldsymbol{\alpha}}(\boldsymbol{x}_0), \tag{1.2}$$

and we notice that problem (1.1)-(1.2) is a standard Markov Decision Process (MDP) with state space $\mathcal{X}^N$, action space A^N, and (randomized) open-loop controls, and is the mathematical framework for reinforcement learning with multiple agents in interaction.

Let us now formulate the asymptotic mean-field problem when the number of agents N goes to infinity. This consists formally in replacing empirical distributions by theoretical ones in the dynamic system and gain functions. The controlled state process X of the representative agent is given by

$$\begin{cases} X_0 &=& \xi_0,\\ X_{t+1} &=& F(X_t^\alpha,\alpha_t,\mathbb{P}^0_{(X_t,\alpha_t)},\varepsilon_{t+1},\varepsilon^0_{t+1}), \quad t\in\mathbb{N}, \end{cases} \tag{1.3}$$

where we have renamed the uniform random sequence $(U_t^1)_{t\in\mathbb{N}}$ and the noise $(\varepsilon_t^1)_{t\in\mathbb{N}}$ by $(U_t)_{t\in\mathbb{N}}$ and $(\varepsilon_t)_{t\in\mathbb{N}}$, and the initial state ξ_0 is a $\mathcal{G}$-measurable random variable, with $\mathcal{G}$ a σ-algebra independent of $(U_t)_{t\in\mathbb{N}}$, $(\varepsilon_t)_{t\in\mathbb{N}}$, $(\varepsilon_t^0)_{t\in\mathbb{N}^\star}$, with distribution law $\mu_0\in\mathcal{P}(\mathcal{X})$ (the set of probability measures on $\mathcal{X}$). The control process α is an A-valued process, adapted with respect to the filtration generated by $\mathcal{G}$, $(U_t)_{t\in\mathbb{N}}$, $(\varepsilon_t)_{t\in\mathbb{N}}$, $(\varepsilon_t^0)_{t\in\mathbb{N}^\star}$, denoted by $\alpha\in\mathcal{A}$. Here $\mathbb{P}^0$ and $\mathbb{E}^0$ represent the conditional probability and expectation knowing the common noise ε^0, and then, given a random variable Y, we denote by $\mathbb{P}^0_Y$ or $\mathcal{L}^0(Y)$ its conditional law knowing ε^0. The McKean-Vlasov (or mean-field) control problem consists in maximizing over randomized open-loop controls α in $\mathcal{A}$ the gain functional

$$V^\alpha(\xi_0) \;:=\; \mathbb{E}\Big[\sum_{t=0}^\infty \beta^t f\big(X_t,\alpha_t,\mathbb{P}^0_{(X_t,\alpha_t)}\big)\Big].$$

The value function to this optimization problem is defined on $\mathcal{P}(\mathcal{X})$ by

$$V(\xi_0) := \sup_{\alpha\in\mathcal{A}} V^\alpha(\xi_0), \tag{1.4}$$

and we recall from [16] that V depends on ξ_0 only through its distribution (invariance in law), and we denote by misuse of notation: $V(\mu_0) = V(\xi_0)$. Problem (1.3)-(1.4) is called mean-field

Markov Decision Process with common noise, or conditional McKean-Vlasov Markov Decision Process (CMKV-MDP in short), with the peculiarity compared to standard MDP coming from the dependence of the state transition on the conditional distribution of the state/action. In view of propagation of chaos for particle systems usually derived for mean-field diffusion process (see [18]), it is expected that CMKV-MDP provides a mean-field approximation of the N-agent MDP model.

While the literature on mean-field control in continuous time, in particular the optimal control of McKean-Vlasov equations, is quite important, see the monograph [5] for an overview and related references, there are rather few papers devoted to the discrete time framework. One of the first works is [11] which studies the convergence of large interacting population process to a simple mean-field model when the state space is finite. The paper [17] studies a discrete-time McKean-Vlasov control problem with feedback controls on finite horizon, and derive the corresponding dynamic programming equation which is explicitly solved in the linear quadratic case. In [6], the authors consider mean-field control on infinite horizon with common noise with a discussion about connections between closed-loop and open-loop policies, and propose Q-learning algortithms. Our companion paper [16] deals with open-loop control and highlights the role of randomized controls with respect to standard Markov Decision Process (MDP). The value function is characterized as a fixed point Bellman equation defined on the space of probability measures, and existence of ϵ-optimal randomized feedback controls is proved. The recent paper [1] studies mean-field control with deterministic closed-loop policies through the lens of MDP theory, and discusses the existence of optimal policies for the limiting mean-field problem as well as for the N-agent problem.

Main contributions. In this paper, we establish a quantitative propagation of result for the N-agent MDP towards the CMKV-MDP. Our contributions are twofold:

1. We show in Theorem 2.1 an explicit rate of convergence of the value functions under some assumptions to be precised later: there exists some positive constant C (depending on the data of the problem) such that for all $\boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N$,

$$\Big|V_N(\boldsymbol{x}) - V\big(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\big)\Big| \quad \leqslant \quad CM_N^\gamma,$$

   where M_N is the mean rate of convergence in Wasserstein distance of the empirical measure (see [9]), and $\gamma \in (0,1]$ is an explicit constant depending on β and F.

2. We prove that any ϵ-optimal randomized feedback policy for the CMKV-MDP (including the case $\epsilon = 0$, i.e., optimal randomized feedback policy whose existence is shown) yields either an approximate optimal feedback control or an approximate randomized feedback control for the N-agent MDP problem, in a constructive sense to be precised later with an explicit rate of convergence, see Theorems 2.2 and 2.3.

While the first statement for convergence of value function is important in theory, the second statement is particularly interesting in practice (but often less studied in the literature) since it means that if the McKean-Vlasov MDP is simpler to solve than the N-agent MDP (some examples and applications to targeted advertising are developed in the PhD thesis [15]), then one can compute an almost optimal randomized feedback policy for the McKean-Vlasov MDP, and then use it in the N-agent MDP: this will guaranty us to have an almost optimal control.

Related literature. The convergence of the N-individual problem to the limiting mean-field control problem has been first rigorously proved in [14] by tightness and martingale arguments for continuous-time controlled McKean-Vlasov equations. This result has been extended in [8] to the common noise case and when there is interaction via the joint distribution of the state and control. The paper [10] proved by viscosity solutions method via the characterization of the Hamilton-Jacobi-Bellman equation the convergence of the value function towards the N-agent problem to the value function of the mean-field control problem in the common noise case but without idiosyncratic noise. Rate of convergence of order $1/N$ has been stated in [12] by Backward Stochastic Differential Equations techniques but under the strong condition that there exists a smooth solution to the Master Bellman equation. The recent paper [4] removed this regularity assumption on the value function, and obtained an algebraic rate of convergence of order $N^{-\gamma}$ for some constant $\gamma \in (0, 1]$. We mention also in the continuous-time framework the paper [7] which derived a rate of convergence of order $N^{-1/2}$ when the state space is finite.

The convergence of the value function in the N-agent problem in a discrete-time mean field framework has been studied in our companion paper [16]. However, it was assumed there that each agent used the same open-loop policy, applied to her own idiosyncratic noise and the common noise. In particular, agent's controls cannot depend upon other agent's idiosyncratic noises, and they have symmetric (or exchangeable) behaviours. This restriction was crucial for using propagation of chaos argument relying on a pathwise comparison between the state and control processes in the N-individual model and the McKean-Vlasov MDPs.

In this paper, we consider that the control of each agent can also depend upon the idiosyncratic noises of all the population, and that they can do so in a completely asymmetric way (i.e. each agent can use a different open-loop policy). This additional flexibility and generality in the definition of controls prevents us from coupling controls between the N-agent and the McKean-Vlasov MDPs in a one-to-one fashion as in [16]. In order to overcome this difficulty, we adopt quite different arguments by coupling the Bellman operators instead of the state/control process of the N-agent and CMKV MDPs. More precisely, the strategy of the proof is the following:
Idea of the proof.

(i) We first derive the Bellman equation for the N-agent MDP, with arguments similar to [16], i.e. we prove that $\mathcal{T}_N V_N = V_N$, where $\mathcal{T}_N$ is the operator defined by

$$\mathcal{T}_N W(\boldsymbol{x}) := \sup_{\mathbf{a} \in A^N} \mathbb{T}_N^{\mathbf{a}} W(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}^N,$$

with

$$\mathbb{T}_N^{\mathbf{a}} W(\boldsymbol{x}) := \frac{1}{N} \sum_{i=1}^N f(x^i, a^i, \frac{1}{N} \sum_{j=1}^N \delta_{(x^j, a^j)}) + \beta \mathbb{E}\big[W\big((F(x^i, a^i, \frac{1}{N} \sum_{j=1}^N \delta_{(x^j, a^j)}, \varepsilon_1^i, \varepsilon_1^0)_{i \in [\![1,N]\!]}\big)\big],$$

for $\boldsymbol{x} = (x^i)_{i \in [\![1,N]\!]} \in \mathcal{X}^N$, $\mathbf{a} = (a^i)_{i \in [\![1,N]\!]} \in A^N$. This property is obtained by seeing the N-agent MDP as a standard MDP on $\mathcal{X}^N$ with actions space A^N.

(ii) Then, we observe that the operators $\mathbb{T}^{\mathrm{a}}$ of the McKean-Vlasov MDP, derived in [16], are, formally, the limits of $\mathbb{T}_N^{\mathbf{a}}$ when $N \to \infty$, for $\mathrm{a} \in L^0(\mathcal{X} \times [0, 1], A)$ and $\mathbf{a} \in A^N$ well coupled. Inspired by this formal observation, we "compare" $\mathbb{T}_N^{\mathbf{a}}$ to $\mathbb{T}^{\mathrm{a}}$ and prove that they are indeed "close" in some sense, for N large. A key point is that $\mathbb{T}_N^{\mathbf{a}}$ is defined on $L_m^\infty(\mathcal{X}^N)$ (the set

of bounded measurable functions on $\mathcal{X}^N$, valued in $\mathbb{R}$) while $\mathbb{T}^a$ is defined on $L^\infty_m(\mathcal{P}(\mathcal{X}))$ (the set of bounded measurable functions on $\mathcal{P}(\mathcal{X})$, valued in $\mathbb{R}$). To compare both type of objects, we introduce a canonical way to associate to a function $W \in L^\infty_m(\mathcal{P}(\mathcal{X}))$ the function $\widetilde{W} \in L^\infty_m(\mathcal{X}^N)$ by setting $\widetilde{W}(\boldsymbol{x}) = W\big(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\big)$.

(iii) Once the proximity between $\mathbb{T}^{\mathbf{a}}_N$ and $\mathbb{T}^a$ is established in a general sense, we prove the proximity of the value functions V_N and V by seeing them as the unique fixed points of the Bellman operators $\mathcal{T}_N = \sup_{\mathbf{a}\in A^N} \mathbb{T}^{\mathbf{a}}_N$ and $\mathcal{T} = \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1],A)} \mathbb{T}^{\mathrm{a}}$, following the intuition that if two contracting operators are close, their unique fixed points should also be close.

(iv) Finally, we provide two procedures to build $\mathcal{O}(\epsilon + M_N^\gamma)$-optimal policies for the N-agent MDP from an ϵ-optimal stationary randomized feedback policy for the McKean-Vlasov MDP. The idea is to view, for each MDP, any ϵ-optimal policy as a policy satisfying the verification theorem, which is a property only linked to the Bellman operator, again following the intuition that if two Bellman operators are close, the policies satisfying their verification results should also be close.

Outline of the paper. The rest of the paper is organized as follows. We state the assumptions and the main results in Section 2, while Section 3 is devoted to their proofs. Finally, we give in Appendix A the proof of existence for optimal randomized feedback policy, and put in Appendix B some results about the Bellman operator for the N-agent MDP problem that are needed in the proof of our convergence results.

2 Main results

2.1 Notations and assumptions

The product space $\mathcal{X} \times A$ is equipped with the metric $\mathbf{d}((x,a),(x',a')) = d(x,x') + d_A(a,a')$, $x, x' \in \mathcal{X}$, $a, a' \in A$. Likewise, we shall endow $\mathcal{X}^N$ with the metric $\boldsymbol{d}_N(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{N}\sum_{i=1}^N d(x^i,x'^i)$ for $\boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]}, \boldsymbol{x}' = (x'^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N$, A^N with the metric $\boldsymbol{d}_{A,N}(\mathbf{a},\mathbf{a}') = \frac{1}{N}\sum_{i=1}^N d_A(a^i,a'^i)$ for $\mathbf{a} = (a^i)_{i\in[\![1,N]\!]}, \mathbf{a}' = (a'^i)_{i\in[\![1,N]\!]} \in A^N$, and $(\mathcal{X}\times A)^N$ with the metric $\mathbf{d}_N((\boldsymbol{x},\mathbf{a}),(\boldsymbol{x}',\mathbf{a}')) = \frac{1}{N}\sum_{i=1}^N \mathbf{d}((x^i,a^i),(x'^i,a'^i))$ for $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}$ and $\mathbf{a},\mathbf{a}' \in A^N$. 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

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

where $\boldsymbol{\Pi}(\mu,\mu')$ is the set of (coupling) probability measures on $\mathcal{Y}\times\mathcal{Y}$ with marginals μ and μ', and we recall the dual Kantorovich-Rubinstein representation

$$\mathcal{W}_d(\mu,\mu') = \sup_{\phi\in Lip_1}\int_{\mathcal{Y}} \phi(y)(\mu-\mu')(\mathrm{d}y), \tag{2.1}$$

where Lip_1 is the set of Lipschitz functions on $\mathcal{Y}$ with Lipschitz constant bounded by 1.

Given $\boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N$, and $\mathbf{a} = (a^i)_{i\in[\![1,N]\!]} \in A^N$, we denote by

$$\mu_{\scriptscriptstyle N}[\boldsymbol{x}] \;:=\; \frac{1}{N}\sum_{i=1}^N \delta_{x^i} \;\in\; \mathcal{P}(\mathcal{X}), \qquad \mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}] \;:=\; \frac{1}{N}\sum_{i=1}^N \delta_{(x^i,a^i)} \;\in\; \mathcal{P}(\mathcal{X}\times A),$$

and we recall that

$$\mathcal{W}_{\mathbf{d}}(\mu_{_N}[\boldsymbol{x},\mathbf{a}],\mu_{_N}[\boldsymbol{x}',\mathbf{a}']) \leqslant \mathbf{d}_N((\boldsymbol{x},\mathbf{a}),(\boldsymbol{x}',\mathbf{a}')). \tag{2.2}$$

Given a random variable Y on $(\Omega,\mathcal{F},\mathbb{P})$, we denote by $\mathbb{P}_Y$ or $\mathcal{L}(Y)$ its distribution law.

We make the following standing assumptions on the state transition function F and on the running reward function f.

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

$$\mathbb{E}\big[d\big(F(x,a,\mu,\varepsilon_1^1,e^0),F(x',a',\mu',\varepsilon_1^1,e^0)\big)\big] \leqslant K_F\big(\mathbf{d}((x,a),(x',a'))+\mathcal{W}_{\mathbf{d}}(\mu,\mu')\big).$$

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

$$|f(x,a,\mu)-f(x',a',\mu')| \leqslant K_f\big(\mathbf{d}((x,a),(x',a'))+\mathcal{W}_{\mathbf{d}}(\mu,\mu')\big).$$

Remark 2.1 We stress the importance of making the regularity assumptions for F in *expectation* only. When $\mathcal{X}$ is finite, F cannot be, strictly speaking, Lipschitz (or even continuous) unless it is constant w.r.t. its mean-field argument (μ and μ' in (**HF$_{\mathbf{lip}}$**)). However, F can be Lipschitz *in expectation*, e.g. once integrated w.r.t. the idiosyncratic noise.

Under Assumption (**HF$_{\mathbf{lip}}$**), we define the constant

$$\gamma \;:=\; \min\big[1,\frac{|\ln\beta|}{\ln(2K_F)_+}\big] \;\in\; (0,1].$$

In the sequel, we denote by $\Delta_{\mathcal{X}}$ (resp. Δ_A and $\Delta_{\mathcal{X}\times A}$) the diameter of the compact metric space $\mathcal{X}$ (resp. A and $\mathcal{X}\times A$), and define

$$M_N := \sup_{\mu\in\mathcal{P}(\mathcal{X}\times A)} \mathbb{E}[\mathcal{W}_{\mathbf{d}}(\mu_N,\mu)], \tag{2.3}$$

where μ_N is the empirical measure $\mu_N=\frac{1}{N}\sum_{n=1}^N\delta_{Y_n}$, $(Y_n)_{1\leqslant n\leqslant N}$ are i.i.d. random variables with law μ. It is know that $M_N \underset{N\to\infty}{\to} 0$, and we recall from [9], and [3] some results about non asymptotic bounds for the mean rate of convergence in Wasserstein distance of the empirical measure.

- 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 $d\geqslant 3$.
- 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 $M_N=\mathcal{O}(N^{-1/\theta})$.

In the sequel C will denote a generic constant that depends only on the data of the problem, namely $\Delta_{\mathcal{X}}$, $\Delta_{\mathcal{X}\times A}$, β, K_F and K_f.

2.2 Convergence of value functions

Our first main result is to quantify the rate of convergence of the value function of the N-agent MDP towards the value function of the CMKV-MDP.

Theorem 2.1 *There exists some positive constant C such that for all $\boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N$,*

$$\Big|V_N(\boldsymbol{x}) - V\big(\mu_{_N}[\boldsymbol{x}]\big)\Big| \quad \leqslant \quad CM_N^{\gamma}.$$

2.3 Approximate optimal policies

Our next results are to show how to obtain approximate optimal control for the N agent MDP from ε-optimal control for CKMV-MDP, and to quantify the accuracy of this approximation.

First, let us recall from [16] the construction of ε-optimal control for CKMV-MDP. The value function V is characterized as the unique fixed point in $L^{\infty}_m(\mathcal{P}(\mathcal{X}))$, the set of bounded measurable real-valued functions on $\mathcal{P}(\mathcal{X})$, of the Bellman equation $V = \mathcal{T}V$, where $\mathcal{T}$ is the Bellman operator defined on $L^{\infty}_m(\mathcal{P}(\mathcal{X}))$ by

$$\begin{aligned} \mathcal{T}W(\mu) &:= \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)} \mathbb{T}^{\mathrm{a}}W(\mu), \\ \text{with} \quad \mathbb{T}^{\mathrm{a}}W(\mu) &:= \mathbb{E}\Big[f(\xi,\mathrm{a}(\xi,U),\mathcal{L}(\xi,\mathrm{a}(\xi,U))) + \beta W\big(\mathbb{P}^0_{F(\xi,\mathrm{a}(\xi,U),\mathcal{L}(\xi,\mathrm{a}(\xi,U)),\varepsilon_1,\varepsilon_1^0)}\big)\Big], \end{aligned} \tag{2.4}$$

for any $(\xi,U)\sim\mu\otimes\mathcal{U}([0,1])$ (it is clear that the right-hand side in (2.4) does not depend on the choice of such (ξ,U)), where $L^0(\mathcal{X}\times[0,1];A)$ is the set of measurable functions from $\mathcal{X}\times[0,1]$ into A. Then, for all $\epsilon>0$, there exists a randomized feedback policy $\mathfrak{a}_\epsilon$, i.e. a measurable function from $\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1]$ into A, denoted by $\mathfrak{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1];A)$, such that for all $\mu\in\mathcal{P}(\mathcal{X})$:

$$V(\mu) - \epsilon \quad \leqslant \quad \mathbb{T}^{\mathfrak{a}_\epsilon(\mu,.)}V(\mu),$$

and we say that $\mathfrak{a}_\epsilon$ is an ϵ-optimal randomized feedback policy for CMKV-MDP. By considering the randomized feedback control $\alpha^\epsilon\in\mathcal{A}$ defined by

$$\alpha^\epsilon_t = \mathfrak{a}_\epsilon(\mathbb{P}^0_{X_t}, X_t, U_t), \quad t\in\mathbb{N}, \tag{2.5}$$

where $(U_t)_{t\in\mathbb{N}}$ is an i.i.d. sequence of random variables, $U_t\sim\mathcal{U}([0,1])$, independent of $\xi_0\sim\mu_0$, and $\boldsymbol{\varepsilon}$, this yields an $O(\epsilon)$-optimal control for $V(\mu_0)$, namely

$$V(\mu_0) - \frac{\epsilon}{1-\beta} \quad \leqslant \quad V^{\alpha^\epsilon}(\xi_0).$$

Actually, we can even take $\epsilon=0$, i.e., get optimal randomized feedback control. The proof for the existence of an optimal randomized feedback policy is inspired by the paper [6], which states the existence of an optimal policy in a closely related model, and is reported in Appendix A.

We now provide two procedures to construct approximate optimal control for the N-agent MDP from an ϵ-optimal randomized feedback policy for CMKV-MDP. The first procedure gives a general approach for getting approximate feedback control for the N-agent MDP.

Theorem 2.2 *Let $\mathfrak{a}_\epsilon$ be an ϵ-optimal randomized feedback policy for CMKV-MDP. Then, there exists a measurable function $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon,N}$ from $\mathcal{X}^N$ into A^N, called feedback policy for the N-agent MDP, such that*

$$\boldsymbol{\pi}^{\mathfrak{a}_\epsilon,N}(\boldsymbol{x}) \in \operatorname*{argmin}_{\mathbf{a}\in A^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}\big(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_{_N}[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)\big), \mu_{_N}[\boldsymbol{x}, \mathbf{a}]\big), \quad \boldsymbol{x} \in \mathcal{X}^N, \tag{2.6}$$

with $(\xi_{\boldsymbol{x}}, U) \sim \mu_{_N}[\boldsymbol{x}] \otimes \mathcal{U}([0,1])$. This yields a feedback control $\boldsymbol{\alpha}^{\epsilon,N} \in \boldsymbol{\mathcal{A}}$ defined by

$$\boldsymbol{\alpha}_t^{\epsilon,N} \quad = \quad \boldsymbol{\pi}^{\mathfrak{a}_\epsilon,N}(\boldsymbol{X}_t), \quad t \in \mathbb{N},$$

which is $O(\varepsilon + M_N^\gamma)$-optimal control for $V_N(\boldsymbol{x}_0)$, namely:

$$V_N(\boldsymbol{x}_0) - C\big[\epsilon + M_N^\gamma\big] \quad \leqslant \quad V_N^{\boldsymbol{\alpha}^{\epsilon,N}}(\boldsymbol{x}_0).$$

Theorem 2.2 provides a generic way to obtain a $\mathcal{O}(\epsilon + M_N^\gamma)$-optimal feedback policy for the N-agent MDP from an ε-optimal randomized feedback policy $\mathfrak{a}_\varepsilon$ for CMKV-MDP, simply by sending actions $\mathbf{a} = (a^i)_{i\in[\![1,N]\!]}$ to the population so that, once in state $\boldsymbol{x}$, the state-action pair $(\boldsymbol{x}, \mathbf{a})$ is empirically distributed as closely as possible to $\mathcal{L}\big(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_{_N}[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)\big)$. However, the computation of this argmin in (2.6) can be difficult in practice.

We propose a second approach which provides a more practical derivation of an approximate optimal control for the N-agent MDP. It will use randomized feedback policy for the N-agent model, defined as a measurable function from $\mathcal{X}^N \times [0,1]^N$ into A^N.

Theorem 2.3 *Let $\mathfrak{a}_\epsilon$ be an ϵ-optimal randomized feedback policy for CMKV-MDP, assumed to satisfy the regularity condition*

$$\mathbb{E}[d_A(\mathfrak{a}_\epsilon(\mu, x, U), \mathfrak{a}_\epsilon(\mu, x', U))] \leqslant K d(x, x'), \quad \forall x, x' \in \mathcal{X},\ \mu \in \mathcal{P}(\mathcal{X}), \tag{2.7}$$

(here $U \sim \mathcal{U}([0,1])$) for some positive constant K. Consider the randomized feedback policy in the N-agent model defined by

$$\boldsymbol{\pi}_r^{\mathfrak{a}_\epsilon,N}(\boldsymbol{x}, \boldsymbol{u}) \quad := \quad \big(\mathfrak{a}_\epsilon(\mu_{_N}[\boldsymbol{x}], x^i, u^i)\big)_{i\in[\![1,N]\!]},$$

for $\boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N$, $\boldsymbol{u} = (u^i)_{i\in[\![1,N]\!]} \in [0,1]^N$. Then, the randomized feedback control $\boldsymbol{\alpha}^{r,\varepsilon,N} \in \boldsymbol{\mathcal{A}}$ defined as

$$\boldsymbol{\alpha}_t^{r,\epsilon,N} \quad = \quad \boldsymbol{\pi}_r^{\mathfrak{a}_\epsilon,N}(\boldsymbol{X}_t, \boldsymbol{U}_t), \quad t \in \mathbb{N},$$

where $\{\boldsymbol{U}_t = (U_t^i)_{i\in[\![1,N]\!]}, t \in \mathbb{N}\}$ is a family of mutually i.i.d. uniform random variables on $[0,1]$, independent of $\mathcal{G}$, $\boldsymbol{\varepsilon} = \big((\varepsilon_t^i)_{i\in[\![1,N]\!]}, \varepsilon_t^0\big)_{t\in\mathbb{N}^\star}$, is an $O(\epsilon + M_N^\gamma)$-optimal control for $V_N(\boldsymbol{x}_0)$, namely:

$$V_N(\boldsymbol{x}_0) - C(1+K)(\varepsilon + M_N^\gamma) \quad \leqslant \quad V_N^{\boldsymbol{\alpha}^{r,\varepsilon,N}}(\boldsymbol{x}_0).$$

Theorem 2.3 provides a simple and natural procedure to get an approximate policy for the N-agent MDP: it corresponds to using an ϵ-optimal randomized feedback policy $\mathfrak{a}_\epsilon$ of the CMKV-MDP, but instead of inputting the theoretical state distribution of the McKean-Vlasov MDP in its mean-field argument, we input the empirical state distribution of the N-agent MDP, and instead of inputting the McKean-Vlasov state in its state argument, we input the N-agent individual states, and moreover, we use a randomization by tossing a coin at any time and for any agent. Notice that the validity of this procedure requires the Lipschitz condition (2.7), which always holds true when the state space $\mathcal{X}$ is finite. Indeed, in this case, the metric on $\mathcal{X}$ is the discrete distance $d(x, x') = 1_{x\neq x'}$, and (2.7) is clearly satisfied with $K = \Delta_A$.

3 Proof of main results

This section is devoted to the proofs of Theorems 2.1, 2.2, and 2.3 about rate of convergence in the propagation of chaos between the N-agent MDP and the limiting conditional McKean-Vlasov MDP. Our approach relies on the Bellman operators of each MDP. By proving their proximity (in a sense to be precised), we will be able to prove on the one hand the proximity of their unique fixed points, hence the convergence of the value functions, and on the other hand that almost optimal randomized feedback policies are directly related to the Bellman operators via the verification result, which will give the convergence of the approximate controls.

3.1 Comparing the Bellman operators

We first introduce the following useful measurable optimal permutation for the coupling of empirical measures.

Definition 3.1 (Measurable optimal permutation) *Let $(\mathcal{Y}, d)$ be a metric space. There exists a measurable map $\sigma : (\boldsymbol{y}, \boldsymbol{y}') \in (\mathcal{Y}^N)^2 \to \sigma^{\boldsymbol{y},\boldsymbol{y}'} \in \mathfrak{S}_N$ (where $\mathfrak{S}_N$ denotes the set of permutations on $[\![1, N]\!]$) such that for all $(\boldsymbol{y}, \boldsymbol{y}') \in (\mathcal{Y}^N)^2$, we have*

$$\mathcal{W}_d\big(\mu_{\scriptscriptstyle N}[\boldsymbol{y}], \mu_{\scriptscriptstyle N}[\boldsymbol{y}']\big) = \boldsymbol{d}_N(\boldsymbol{y}, \boldsymbol{y}'_{\sigma^{\boldsymbol{y},\boldsymbol{y}'}}), \tag{3.1}$$

where we set $\boldsymbol{y}'_{\sigma^{\boldsymbol{y},\boldsymbol{y}'}} = (y'^{\sigma^{\boldsymbol{y},\boldsymbol{y}'}}_i)_{i\in[\![1,N]\!]}$ for $\boldsymbol{y}' = (y'^i)_{i\in[\![1,N]\!]}$.

Proof. It is a well known result (see [19]) that, given $(\boldsymbol{y}, \boldsymbol{y}') \in (\mathcal{Y}^N)^2$, there exists a permutation $\sigma^{\boldsymbol{y},\boldsymbol{y}'} \in \mathfrak{S}_N$ realizing an optimal coupling between $\mu_N[\boldsymbol{y}]$, $\mu_{\scriptscriptstyle N}[\boldsymbol{y}'] \in \mathcal{P}(\mathcal{Y})$, i.e., s.t. (3.1) holds. Let us check that this optimal permutation can be represented as a measurable function of $(\boldsymbol{y}, \boldsymbol{y}') \in (\mathcal{Y}^N)^2$. Let $n \in [\![1, N!]\!] \mapsto \sigma^n \in \mathfrak{S}_N$ be some bijection. Notice that the function

$$\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}^N \mapsto \big(\boldsymbol{d}_N(\boldsymbol{y}, \boldsymbol{y}'_{\sigma^n})\big)_{n\in[\![1,N!]\!]} \in \mathbb{R}^{N!}$$

is continuous, hence measurable. Furthermore, it is clear that the function

$$\boldsymbol{z} \in \mathbb{R}^{N!} \mapsto \min\big[\operatorname*{argmin}_{n\in N!} z^n)\big]$$

is measurable. Denoting by

$$n_{min}(\boldsymbol{y}, \boldsymbol{y}') \quad := \quad \min\big[\operatorname*{argmin}_{n\in N!} \boldsymbol{d}_N(\boldsymbol{y}, \boldsymbol{y}'_{\sigma^n})\big],$$

it follows that the function $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{X}^N \mapsto \sigma^{\boldsymbol{y},\boldsymbol{y}'} = \sigma^{n_{min}(\boldsymbol{y},\boldsymbol{y}')}$ is a measurable representation of the optimal permutation. □

We now study the "proximity" between the Bellman operator of the CMKV-MDP given in (2.4), and the Bellman operator of the N-agent problem, viewed as a MDP with state space $\mathcal{X}^N$, action space A^N, noise sequence $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_t)_{t\in\mathbb{N}^\star}$ with $\boldsymbol{\varepsilon}_t := ((\varepsilon^i_t)_{i\in[\![1,N]\!]}, \varepsilon^0_t)$ valued in $E^N \times E^0$, state transition function

$$\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, \mathbf{e}) \quad := \quad \Big(F(x^i, a^i, \mu_{\scriptscriptstyle N}[\boldsymbol{x}, \mathbf{a}], e^i, e^0)\Big)_{i\in[\![1,N]\!]}, \quad \mathbf{e} = ((e^i)_{i\in[\![1,N]\!]}, e^0) \in E^N \times E^0,$$

and reward function

$$\boldsymbol{f}(\boldsymbol{x},\mathbf{a}) \;=\; \frac{1}{N}\sum_{i=1}^{N} f\big(x^i,a^i,\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}]\big), \quad \boldsymbol{x}=(x^i)_{i\in[\![1,N]\!]},\ \mathbf{a}=(a^i)_{i\in[\![1,N]\!]}.$$

Denoting by $L^\infty_m(\mathcal{X}^N)$ the subset of measurable functions in $L^\infty(\mathcal{X}^N)$ (the set of bounded real-valued functions on $\mathcal{X}^N$), the Bellman "operator" $\mathcal{T}_N : L^\infty_m(\mathcal{X}^N) \to L^\infty(\mathcal{X}^N)$ of the N-agent MDP is defined for any $W \in L^\infty_m(\mathcal{X}^N)$ by:

$$\mathcal{T}_N W(\boldsymbol{x}) := \sup_{\mathbf{a}\in A^N} \mathbb{T}^{\mathbf{a}}_N W(\boldsymbol{x}), \quad \boldsymbol{x}\in\mathcal{X}^N,$$

where

$$\mathbb{T}^{\mathbf{a}}_N W(\boldsymbol{x}) := \ \boldsymbol{f}(\boldsymbol{x},\mathbf{a}) + \beta\mathbb{E}\big[W\big(\boldsymbol{F}(\boldsymbol{x},\mathbf{a},\boldsymbol{\varepsilon}_1)\big)\big], \quad \boldsymbol{x}\in\mathcal{X}^N,\ \mathbf{a}\in A^N.$$

The characterization of the value function V_N and optimal controls for the N-agent MDP via the Bellman operator $\mathcal{T}_N$ is stated in Appendix B.

We aim to quantify how "close" $\mathbb{T}^{\mathbf{a}}_N$ and $\mathbb{T}^{\mathrm{a}}$ are when $\mathbf{a}$ and a are close in a sense to be precised. Notice that the N-agent operator $\mathbb{T}^{\mathbf{a}}_N$ is defined on $L^\infty_m(\mathcal{X}^N)$ while the McKean-Vlasov operator $\mathbb{T}^{\mathrm{a}}$ is defined on $L^\infty_m(\mathcal{P}(\mathcal{X}))$. There is however a natural way to compare them by means of an "unlifting" procedure. To any function $W \in L^\infty_m(\mathcal{P}(\mathcal{X}))$, we associate the unlifted function $\widetilde{W} \in L^\infty_m(\mathcal{X}^N)$ defined by

$$\widetilde{W}(\boldsymbol{x}) := W(\mu_{\scriptscriptstyle N}[\boldsymbol{x}]), \quad \forall \boldsymbol{x}\in\mathcal{X}^N.$$

We recall from [16] that the value function V of the CMKV-MDP is γ-Hölder:

$$|V(\mu)-V(\mu')| \leqslant\ K_\star\big(\mathcal{W}_d(\mu,\mu')\big)^\gamma, \quad \forall \mu,\mu'\in\mathcal{P}(\mathcal{X}), \tag{3.2}$$

for some constant $K_\star$ depending on K_F, β and $\Delta_{\mathcal{X}}$.

Lemma 3.1 *There exists some positive constant C such that for all $\mathrm{a} \in L^0(\mathcal{X}\times[0,1];A)$, $\mathbf{a}\in A^N$, $\boldsymbol{x}\in\mathcal{X}^N$ and $(\xi_{\boldsymbol{x}},U)\sim\mu_{\scriptscriptstyle N}[\boldsymbol{x}]\otimes\mathcal{U}([0,1])$,*

$$|\widetilde{\mathbb{T}^{\mathrm{a}}V}(\boldsymbol{x}) - \mathbb{T}^{\mathbf{a}}_N\breve{V}(\boldsymbol{x})| \quad\leqslant\quad C\Big[\big(\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}]\big)\big)^\gamma + M_N^\gamma\Big].$$

Proof. For any $\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)$, and $\mathbf{a}=(a^i)_{i\in[\![1,N]\!]}\in A^N$, we have

$$\begin{aligned}
&\widetilde{\mathbb{T}^{\mathrm{a}}V}(\boldsymbol{x}) - \mathbb{T}^{\mathbf{a}}_N\breve{V}(\boldsymbol{x})\\
&=\ \mathbb{E}\Big[f(\xi_{\boldsymbol{x}},a(\xi_{\boldsymbol{x}},U),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U))) - \frac{1}{N}\sum_{i=1}^N f(x^i,a^i,\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}])\Big]\\
&\quad+\ \beta\mathbb{E}\Big[V\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\varepsilon_1,\varepsilon_1^0)}\big) - V\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,a^i,\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)}\big)\Big].
\end{aligned} \tag{3.3}$$

We write

$$\mathbb{E}\Big[f(\xi_{\boldsymbol{x}},a(\xi_{\boldsymbol{x}},U),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)))\Big] - \frac{1}{N}\sum_{i=1}^N f(x^i,\mathbf{a}^i,\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}]) \quad=\quad \hat{f}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U))) - \hat{f}(\mu_{\scriptscriptstyle N}[\boldsymbol{x},\mathbf{a}]),$$

where $\hat{f}(\mu) = \int f(x', a', \mu)\mu(\mathrm{d}x', \mathrm{d}a')$ for all $\mu \in \mathcal{P}(\mathcal{X} \times A)$. Notice that for $\mu, \mu' \in \mathcal{P}(\mathcal{X} \times A)$, we have

$$\begin{aligned}\hat{f}(\mu) - \hat{f}(\mu') &= \int f(x', a', \mu)(\mu - \mu')(\mathrm{d}x', \mathrm{d}a') + \int (f(x', a', \mu) - f(x', a', \mu'))\mu'(\mathrm{d}x', \mathrm{d}a') \\ &\leqslant K_f \mathcal{W}_{\mathbf{d}}(\mu, \mu') + K_f \mathcal{W}_{\mathbf{d}}(\mu, \mu') = 2K_f \mathcal{W}_{\mathbf{d}}(\mu, \mu'),\end{aligned}$$

from the Kantorovich-Rubinstein dual representation (2.1) and ($\mathbf{Hf_{lip}}$). It follows that

$$\begin{aligned}&\Big|\mathbb{E}\Big[f\big(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U))\big)\Big] - \frac{1}{N}\sum_{i=1}^N f(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}])\Big| \qquad (3.4) \\ &\leqslant 2K_f \mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]).\end{aligned}$$

Let us next focus on the second term in (3.3). As V is γ-Hölder with constant factor $K_\star$, we have

$$\begin{aligned}&\Big|\mathbb{E}\Big[V\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \varepsilon_1, \varepsilon_1^0)}\big) - V\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}], \varepsilon_1^i, \varepsilon_1^0)}\big)\Big]\Big| \\ &\leqslant K_\star \mathbb{E}\Big[\mathcal{W}_d\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \varepsilon_1^i, \varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}], \varepsilon_1^i, \varepsilon_1^0)}\big)\Big]^\gamma, \qquad (3.5)\end{aligned}$$

by Jensen's inequality. Let $(\xi^i, U_0^i)_{i \in [\![1,N]\!]}$ be N i.i.d. random variables, independent of $\boldsymbol{\varepsilon}_1$, such that $(\xi^i, U_0^i) \sim \mu_N[\boldsymbol{x}] \otimes \mathcal{U}([0,1])$, $i \in [\![1, N]\!]$. For any i.i.d. random variables $(\tilde{\varepsilon}_1^i)_{i \in [\![1,N]\!]}$ such that

$$((\xi^i, U_0^i, \tilde{\varepsilon}_1^i)_{i \in [\![1,N]\!]}, \varepsilon_1^0) \stackrel{d}{=} ((\xi^i, U_0^i, \varepsilon_1^i)_{i \in [\![1,N]\!]}, \varepsilon_1^0), \qquad (3.6)$$

we have

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{W}_d\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \varepsilon_1^i, \varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}], \varepsilon_1^i, \varepsilon_1^0)}\big)\Big] \\ &\leqslant \mathbb{E}\Big[\mathcal{W}_d\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \varepsilon_1^i, \varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(\xi^i, \mathrm{a}(\xi^i, U_0^i), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \tilde{\varepsilon}_1^i, \varepsilon_1^0)}\big)\Big] \\ &+ \mathbb{E}\Big[\mathcal{W}_d\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(\xi^i, \mathrm{a}(\xi^i, U_0^i), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \tilde{\varepsilon}_1^i, \varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}], \varepsilon_1^i, \varepsilon_1^0)}\big)\Big] \\ &\leqslant M_N + \mathbb{E}\Big[\mathcal{W}_d\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(\xi^i, \mathrm{a}(\xi^i, U_0^i), \mathcal{L}(\xi_{\boldsymbol{x}}, \mathrm{a}(\xi_{\boldsymbol{x}}, U)), \tilde{\varepsilon}_1^i, \varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i, a^i, \mu_N[\boldsymbol{x}, \mathbf{a}], \varepsilon_1^i, \varepsilon_1^0)}\big)\Big], \qquad (3.7)\end{aligned}$$

by definition of M_N in (2.3). Let us now consider the random permutation $\sigma^{(\xi^i, \mathrm{a}(\xi^i, U_0^i))_{i \in [\![1,N]\!]}, (x^i, a^i)_{i \in [\![1,N]\!]}}$ defined in Definition 3.1 that we shall, to simplify notations, simply denote by σ. Notice that as $(\xi^i, \mathrm{a}(\xi^i, U_0^i))_{i \in [\![1,N]\!]} \perp\!\!\!\perp (\varepsilon_1^i)_{i \in [\![1,N]\!]}$, we clearly see that $(\tilde{\varepsilon}_1^i)_{i \in [\![1,N]\!]} := (\varepsilon_1^{(\sigma^{-1})_i})_{i \in [\![1,N]\!]}$ satisfies the required condition (3.6). Therefore the above relation applies to $(\tilde{\varepsilon}_1^i)_{i \in [\![1,N]\!]} = (\varepsilon_1^{(\sigma^{-1})_i})_{i \in [\![1,N]\!]}$.

For such $(\tilde{\varepsilon}_1^i)_{i\in[\![1,N]\!]}$, we get

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{W}_d\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(\xi^i,\mathrm{a}(\xi^i,U_0^i),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\varepsilon_1^{(\sigma^{-1})_i},\varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,a^i,\mu_N[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)}\big)\Big]\\
= \;& \mathbb{E}\Big[\mathcal{W}_d\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(\xi^{\sigma_i},\mathrm{a}(\xi^{\sigma_i},U_0^{\sigma_i}),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\varepsilon_1^i,\varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,a^i,\mu_N[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)}\big)\Big]\\
\leqslant \;& \frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big[d\big(F(\xi^{\sigma_i},\mathrm{a}(\xi^{\sigma_i},U_0^{\sigma_i}),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\varepsilon_1^i,\varepsilon_1^0), F(x^i,a^i,\mu_N[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)\big)\Big]\\
\leqslant \;& K_F\frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big[\mathbf{d}\big((\xi^{\sigma_i},\mathrm{a}(\xi^{\sigma_i},U_0^{\sigma_i})),(x^i,a^i)\big) + \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}]\big)\Big]\\
= \;& K_F\mathbb{E}[\mathcal{W}_{\mathbf{d}}(\frac{1}{N}\sum_{i=1}^N \delta_{(\xi^i,\mathrm{a}(\xi^i,U_0^i))},\mu_N[\boldsymbol{x},\mathbf{a}]) + \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}]\big)\Big]\\
\leqslant \;& K_F\Big(M_N + 2\mathbb{E}\big[\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\big]\Big),
\end{aligned}$$

where the first inequality comes from (2.2), the second one is derived by conditioning w.r.t. $((\xi^i,U_0^i)_{i\in[\![1,N]\!]},\varepsilon_1^0)$ and using the regularity in expectation of F in ($\mathbf{HF_{lip}}$), the last equality holds true by definition of the permutation σ realizing the optimal coupling (3.1), and the last inequality from the definition of M_N. Recalling (3.7), we then have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{W}_d\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U),\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\varepsilon_1^i,\varepsilon_1^0)}, \frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,a^i,\mu_N[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)}\big)\Big]\\
\leqslant \;& (1+K_F)M_N + 2K_F\mathbb{E}\big[\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\big]
\end{aligned}$$

which implies by (3.5)

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\mathbb{P}^0_{F(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U),\mathcal{L}(\xi,\mathrm{a}(\xi,U)),\varepsilon_1,\varepsilon_1^0)}\big) - V\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,a^i,\mu_N[\boldsymbol{x},\mathbf{a}],\varepsilon_1^i,\varepsilon_1^0)}\big)\Big]\\
\leqslant \;& K_\star\Big((1+K_F)M_N + 2K_F\mathbb{E}\big[\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\big]\Big)^\gamma.
\end{aligned}$$

Together with (3.4), and plugging into (3.3), we obtain finally

$$\begin{aligned}
&\Big|\widetilde{\mathbb{T}^a V}(\boldsymbol{x}) - \mathbb{T}_N^{\mathbf{a}}\breve{V}(\boldsymbol{x})\Big|\\
\leqslant \;& 2K_f\mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\Big] + K_\star\Big((1+K_F)M_N + 2K_F\mathbb{E}\big[\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\big]\Big)^\gamma\\
\leqslant \;& C\Big\{\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}]) + \Big(\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])\Big)^\gamma + M_N^\gamma\Big\}
\end{aligned}$$

(recall that $\gamma \leqslant 1$), for some constant C depending only on $K_\star$, K_f, K_F, where we also use the fact that $\mathcal{W}_{\mathbf{d}}(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_N[\boldsymbol{x},\mathbf{a}])$ is bounded by a constant depending on the diameter of the compact set $\mathcal{X}\times A$. This ends the proof. □

3.2 Proof of Theorem 2.1

Lemma 3.1 means that given $\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)$, $\mathbf{a}\in A^N$, and for $\boldsymbol{x}\in\mathcal{X}^N$, the Wasserstein distance between the distribution law of $(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U))$ (where $(\xi_{\boldsymbol{x}},U)\sim\mu_N[\boldsymbol{x}]\otimes\mathcal{U}([0,1])$), and

the empirical measure $\mu_{_N}[\boldsymbol{x},\mathbf{a}]$ is small (and N large), then $\mathbb{T}^a V \simeq \mathbb{T}^{\mathbf{a}}_N \breve{V}$. It is thus natural to look for suitable choices of $\mathrm{a} \in L^0(\mathcal{X}\times[0,1];A)$, $\mathbf{a}\in A^N$ so that the above Wasserstein distance is as small as possible. This is quantified in the following result.

Lemma 3.2 *Fix $\boldsymbol{x}\in\mathcal{X}^N$. Then, for any $\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)$, there exists $\mathbf{a}^{\mathrm{a}}\in A^N$ such that*

$$\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\mathbf{a}^{\mathrm{a}}]\big) \;\leqslant\; 2M_N,$$

where $(\xi_{\boldsymbol{x}},U)\sim\mu_{_N}[\boldsymbol{x}]\otimes\mathcal{U}([0,1])$. Conversely, for any $\mathbf{a}\in A^N$, there exists $\mathrm{a}^{\mathbf{a}}\in L^0(\mathcal{X}\times[0,1];A)$ such that

$$\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}^{\mathbf{a}}(\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\mathbf{a}]\big) \;=\; 0.$$

Proof. Fix $\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)$. Let us consider $\boldsymbol{\xi}=(\xi^i)_{i\in[\![1,N]\!]}$ i.i.d. with common distribution $\mu_{_N}[\boldsymbol{x}]$, independent from $\boldsymbol{U}_0=(U_0^i)_{i\in[\![1,N]\!]}$ i.i.d. $\sim\mathcal{U}([0,1])$. We have

$$\begin{aligned}
&\mathbb{E}\big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N\delta_{x^i,\mathrm{a}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)}\big)\big]\\
\leqslant\;&\mathbb{E}\big[\mathcal{W}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N\delta_{\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},\mathrm{a}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)}\big)+\mathcal{W}_{\mathbf{d}}\big(\frac{1}{N}\sum_{i=1}^N\delta_{\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},\mathrm{a}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)},\frac{1}{N}\sum_{i=1}^N\delta_{x^i,\mathrm{a}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)}\big)\big]\\
\leqslant\;&M_N+\mathbb{E}\big[\frac{1}{N}\sum_{i=1}^N d(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},x^i)\big]\;\leqslant\;2M_N,
\end{aligned}$$

where we used the definition of M_N and (2.2) in the second inequality, and definition of $\sigma^{\boldsymbol{\xi},\boldsymbol{x}}$ in the last inequality. It follows that

$$\mathbb{P}\big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N\delta_{x^i,\mathrm{a}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)}\big)\leqslant 2M_N\big] \;>\; 0,$$

which implies that there exists a vector $\mathbf{a}\in A^N$ such that

$$\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}(\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\mathbf{a}]\big) \;\leqslant\; 2M_N.$$

On the other hand, given such an $\mathbf{a}\in A^N$, there clearly exists $\mathrm{a}^{\mathbf{a}}\in L^0(\mathcal{X}\times[0,1];A)$ such that $\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}^{\mathbf{a}}(\xi_{\boldsymbol{x}},U))=\mu_{_N}[\boldsymbol{x},\mathbf{a}]$: indeed, by considering $(\tilde\xi,\tilde\alpha)\sim\mu_{_N}[\boldsymbol{x},\mathbf{a}]$, it suffices to choose $\mathrm{a}^{\mathbf{a}}$ as a kernel for simulating the conditional distribution of $\tilde\alpha$ knowing $\tilde\xi$. We then have

$$\mathcal{W}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}^{\mathbf{a}}(\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\mathbf{a}]\big) \;=\; 0.$$

□

By combining the general comparison of Bellman operators in Lemma 3.1 with the coupling result in Lemma 3.2, we can now prove the propagation of chaos of value functions.

Proof of Theorem 2.1. From the fixed point equation for V with Bellman operator $\mathcal{T}$ in (2.4), we have

$$\begin{aligned}
\breve{V}(\boldsymbol{x}) &= \widetilde{\mathcal{T}V}(\boldsymbol{x})\\
&= \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)}\widetilde{\mathbb{T}^aV}(\boldsymbol{x}) \;\leqslant\; \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)}\mathbb{T}^{\mathbf{a}^{\mathrm{a}}}_N\breve{V}(\boldsymbol{x})+CM_N^\gamma\\
&\leqslant \mathcal{T}_N\breve{V}(\boldsymbol{x})+CM_N^\gamma,
\end{aligned}$$

where we used Lemma 3.1 and Lemma 3.2 in the first inequality, and the definition of $\mathcal{T}_N$ in the last one. Since V_N is a fixed point of $\mathcal{T}_N$ (see Proposition B.2), we then have:

$$(\check{V} - V_N)(\boldsymbol{x}) \;\leqslant\; (\mathcal{T}_N\check{V} - \mathcal{T}_N V_N)(\boldsymbol{x}) + CM_N^\gamma,$$

and thus by definition of $\mathcal{T}_N$,

$$(\check{V} - V_N)(\boldsymbol{x}) \;\leqslant\; \beta \sup_{\boldsymbol{x}' \in \mathcal{X}^N} (\check{V} - V_N)(\boldsymbol{x}') + CM_N^\gamma,$$

which implies

$$\sup_{\boldsymbol{x} \in \mathcal{X}^N} (\check{V} - V_N)(\boldsymbol{x}) \;\leqslant\; CM_N^\gamma.$$

Likewise, by Lemma 3.1 and Lemma 3.2, we have

$$\begin{aligned}\check{V}(\boldsymbol{x}) \;&=\; \widetilde{\mathcal{T}V}(\boldsymbol{x}) \;=\; \sup_{\mathrm{a} \in L^0(\mathcal{X}\times[0,1];A)} \widetilde{\mathbb{T}^{\mathrm{a}}V}(\boldsymbol{x}) \;\geqslant\; \sup_{\mathbf{a} \in A^N} \widetilde{\mathbb{T}^{\mathbf{a}}V}(x) \\ &\geqslant\; \sup_{\mathbf{a} \in A^N} \mathbb{T}_N^{\mathbf{a}}\check{V}(\boldsymbol{x}) - CM_N^\gamma \;=\; \mathcal{T}_N\check{V}(\boldsymbol{x}) - CM_N^\gamma,\end{aligned}$$

and using the fact that V_N is a fixed point of $\mathcal{T}_N$, we obtain similarly

$$\sup_{\boldsymbol{x} \in \mathcal{X}^N} (V_N - \check{V})(\boldsymbol{x}) \;\leqslant\; CM_N^\gamma,$$

which concludes the proof.

3.3 Proof of Theorem 2.2

We start with a general result estimating the efficiency of a feedback policy for the N-agent MDP by "comparing" it to an ϵ-optimal randomized feedback policy for the CMKV- MDP.

Lemma 3.3 *Let $\mathfrak{a}_\epsilon$ be an ϵ-optimal randomized feedback policy for the CMKV-MDP, and $\mathbf{a} \in A^N$. Then, there exists some positive constant C (depending only on $\Delta_{\mathcal{X}\times A}$, β, K_F, K_f) such that for all $\boldsymbol{x} \in \mathcal{X}^N$,*

$$\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x}) \;\geqslant\; V_N(\boldsymbol{x}) - \epsilon - C\big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]\big)^\gamma + M_N^\gamma\big].$$

Proof. Fix $\boldsymbol{x} \in \mathcal{X}^N$, $\mathbf{a} \in A^N$, and define $\mathrm{a}_\epsilon \in L^0(\mathcal{X} \times [0,1]; A)$ by $\mathrm{a}_\epsilon(x,u) = \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], x, u)$ for $x \in \mathcal{X}$, $u \in [0,1]$. By Theorem 2.1 and the β-contracting property of $\mathbb{T}_N^{\mathbf{a}}$, we have

$$|\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x}) - \mathbb{T}_N^{\mathbf{a}}\check{V}(\boldsymbol{x})| \;\leqslant\; \beta\|V_N(\boldsymbol{x}) - \check{V}(\boldsymbol{x})\|_{\mathcal{X}^N} \;\leqslant\; CM_N^\gamma,$$

and so

$$\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x}) \;\geqslant\; \mathbb{T}_N^{\mathbf{a}}\check{V}(\boldsymbol{x}) - CM_N^\gamma.$$

Together with Lemma 3.1, this yields

$$\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x}) \geqslant\; \widetilde{\mathbb{T}^{\mathrm{a}_\epsilon}V}(\boldsymbol{x}) - C\big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi, \mathrm{a}_\epsilon(\xi, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]\big)^\gamma + M_N^\gamma\big]. \tag{3.8}$$

Denote by α^ϵ the randomized feedback control associated via (2.5) to the randomized feedback policy $\mathfrak{a}_\epsilon$. Then, notice that the gain functional $V^{\alpha^\epsilon}(\xi)$ depends on ξ only through its law $\mu = \mathcal{L}(\xi)$, and we set $V^{\alpha^\epsilon}(\mu) = V^{\alpha^\epsilon}(\xi)$ when $\xi \sim \mu$. Since $V \geqslant V^{\alpha^\epsilon}$, and by the monotonicity of $\mathbb{T}^{\mathrm{a}_\epsilon}$, we have

$$\mathbb{T}^{\mathrm{a}_\epsilon} V(\mu_N[\boldsymbol{x}]) \;\geqslant\; \mathbb{T}^{\mathrm{a}_\epsilon} V^{\alpha_\epsilon}(\mu_N[\boldsymbol{x}]) \;=\; V^{\alpha^\epsilon}(\mu_N[\boldsymbol{x}]) \;\geqslant\; V(\mu_N[\boldsymbol{x}]) - C\epsilon$$

by recalling that V^{α^ϵ} is a fixed point of $\mathbb{T}^{\mathrm{a}_\epsilon}$, and using the fact that $\mathfrak{a}_\epsilon$ is an ϵ-optimal randomized feedback policy for the CMKV-MDP. From Theorem 2.1, this implies that

$$\widetilde{\mathbb{T}^{\mathrm{a}_\epsilon} V}(\boldsymbol{x}) \;\geqslant\; V_N(\boldsymbol{x}) - C\big[\epsilon + M_N^\gamma\big],$$

which proved the required result when combined with (3.8). □

Let us denote by $L^0(\mathcal{X}^N; A^N)$ the set of measurable functions from $\mathcal{X}^N$ into A^N. Given a feedback policy $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; A^N)$ for the N-agent problem, the associated feedback control is the unique control $\boldsymbol{\alpha}^{\boldsymbol{\pi}}$ defined by $\boldsymbol{\alpha}^{\boldsymbol{\pi}}_t = \boldsymbol{\pi}(\mathbf{X}_t)$, $t \in \mathbb{N}$. By misuse of notation, we denote $V_N^{\boldsymbol{\pi}} = V_N^{\boldsymbol{\alpha}^{\boldsymbol{\pi}}}$. Let us then introduce the operator $\mathcal{T}_N^{\boldsymbol{\pi}}$ on $L_m^\infty(\mathcal{X}^N)$, defined by

$$\mathcal{T}_N^{\boldsymbol{\pi}} W(\boldsymbol{x}) = \; \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})) + \beta \mathbb{E}\big[W\big(\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\varepsilon}_1)\big)\big], \quad \boldsymbol{x} \in \mathcal{X}^N.$$

Proposition 3.1 *Let $\mathfrak{a}_\epsilon$ be an ϵ-optimal randomized feedback policy for the CMKV-MDP, and consider any feedback policy $\boldsymbol{\pi}$ for the N-agent MDP. Then, the feedback control $\boldsymbol{\alpha}^{\boldsymbol{\pi}}$ is*

$$\mathcal{O}(\epsilon + \sup_{\boldsymbol{x} \in \mathcal{X}^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})]\big)^\gamma + M_N^\gamma)\text{-optimal for } V_N(\boldsymbol{x}_0),$$

where $(\xi_{\boldsymbol{x}}, U) \sim \mu_N[\boldsymbol{x}] \otimes \mathcal{U}([0,1])$, namely

$$V_N(\boldsymbol{x}_0) - C\big[\epsilon + \sup_{\boldsymbol{x} \in \mathcal{X}^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x})]\big)^\gamma + M_N^\gamma\big] \;\leqslant\; V_N^{\boldsymbol{\pi}}(\boldsymbol{x}_0).$$

Proof. Fix $\boldsymbol{x} \in \mathcal{X}^N$, and let $\mathbf{a} = \boldsymbol{\pi}(\boldsymbol{x}) \in A^N$. By definition, we have $\mathcal{T}_N^{\boldsymbol{\pi}} V_N(\boldsymbol{x}) = \mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x})$. By Lemma 3.3, we thus have

$$\mathcal{T}_N^{\boldsymbol{\pi}} V_N(\boldsymbol{x}) \;\geqslant\; V_N(\boldsymbol{x}) - \epsilon - C\big[\sup_{\boldsymbol{x} \in \mathcal{X}^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]\big)^\gamma + M_N^\gamma\big],$$

and we conclude by the verification result in Lemma B.5. □

Proposition 3.1 has an important implication: it means that a feedback policy $\boldsymbol{\pi}$ for the N-agent MDP yields the better performance whenever it assigns for each state $\boldsymbol{x}$ the action $\boldsymbol{\pi}(\boldsymbol{x})$ that achieves the minimum of

$$\mathbf{a} \in A^N \;\mapsto\; \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\varepsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]\big).$$

Let us check that one can choose a measurable version of this argmin.

Lemma 3.4 *Let $\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$. Then, there exists a measurable function $\boldsymbol{\pi}^\star : \mathcal{X}^N \to A^N$ such that*

$$\boldsymbol{\pi}^\star(\boldsymbol{x}) \;\in\; \operatorname*{argmin}_{\mathbf{a} \in A^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \mathbf{a}]\big), \quad \boldsymbol{x} \in \mathcal{X}^N.$$

Proof. Notice that the function

$$h(\boldsymbol{x}, \mathbf{a}) \quad := \quad \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}(\mu_{_N}[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_{_N}[\boldsymbol{x}, \mathbf{a}]\big)$$

is such that for $\mathbf{a} \in A^N$, $h(\cdot, \mathbf{a})$ is measurable, and for $\boldsymbol{x} \in \mathcal{X}^N$, $h(\boldsymbol{x}, \cdot)$ is continuous. Let us then show that one can measurably select $\operatorname{argmin}_{\mathbf{a} \in A^N} h(\boldsymbol{x}, \mathbf{a})$ w.r.t. $\boldsymbol{x}$. Consider a dense sequence $(\mathbf{a}_n)_{n \in \mathbb{N}} \subset A^N$ (its existence is guaranteed by the fact that A^N is a compact metric space), and define by recursion the sequence of measurable functions $\boldsymbol{\pi}_n : \mathcal{X}^N \to A^N$ as

$$\begin{aligned} \boldsymbol{\pi}_0(\boldsymbol{x}) &= \mathbf{a}_0 \\ \boldsymbol{\pi}_{n+1}(\boldsymbol{x}) &= \begin{cases} \boldsymbol{\pi}_n(\boldsymbol{x}) \text{ if } h(\boldsymbol{x}, \boldsymbol{\pi}_n(\boldsymbol{x})) \leqslant h(\boldsymbol{x}, \mathbf{a}_{n+1}) \\ \mathbf{a}_{n+1} \text{ else.} \end{cases} \end{aligned}$$

The measurability of $\boldsymbol{\pi}_n$ is easily established by induction on n: For $n = 0$, it is clear. Assuming that $\boldsymbol{\pi}_n$ is measurable, and denoting

$$g_n(\boldsymbol{x}) \quad = \quad h(\boldsymbol{x}, \boldsymbol{\pi}_n(\boldsymbol{x})) - h(\boldsymbol{x}, \mathbf{a}_{n+1}), \forall \boldsymbol{x} \in \mathcal{X}^N,$$

notice that for any measurable set $B \subset A^N$, we have

$$[\boldsymbol{\pi}_{n+1}]^{-1}(B) = \begin{cases} [\boldsymbol{\pi}_n]^{-1}(B) \cap g_n^{-1}(\mathbb{R}_-) \text{ if } \mathbf{a}_{n+1} \notin B, \\ \big([\boldsymbol{\pi}_n]^{-1}(B) \cap g_n^{-1}(\mathbb{R}_-)\big) \cup g_n^{-1}(\mathbb{R}_+^\star) \text{ if } \mathbf{a}_{n+1} \in B, \end{cases}$$

which is clearly a measurable set, and proves the induction. Then, let us consider an embedding $\phi : A^N \to [0,1]$ such that ϕ and ϕ^{-1} are uniformly continuous (see Lemma C.2 in [16]). Then, $(\phi \circ \boldsymbol{\pi}_n)_{n \in \mathbb{N}}$ denotes a sequence of measurable functions from $\mathcal{X}^N$ to $\phi(A^N) \subset [0,1]$. It is well known that the function $\liminf_{n \in \mathbb{N}}(\phi \circ \boldsymbol{\pi}_n)$ is then measurable from $\mathcal{X}^N$ to $\phi(A^N)$ (we here use the fact that ϕ is continuous and $\phi(A^N)$ is closed, which ensures that the lim inf takes its values in $\phi(A^N)$). Finally, let us denote $\boldsymbol{\pi}^\star : \mathcal{X}^N \to A^N$ defined by

$$\boldsymbol{\pi}^\star \quad = \quad \phi^{-1} \circ \liminf_{n \in \mathbb{N}} (\phi \circ \boldsymbol{\pi}_n).$$

$\boldsymbol{\pi}^\star$ is then measurable by composition. Furthermore, for any $\boldsymbol{x} \in \mathcal{X}^N$, $\phi \circ \boldsymbol{\pi}^\star(\boldsymbol{x}) = \liminf_{n \in \mathbb{N}}(\phi \circ \boldsymbol{\pi}_n(\boldsymbol{x}))$ is an accumulation point of the sequence $(\phi \circ \boldsymbol{\pi}_n(\boldsymbol{x}))_{n \in \mathbb{N}}$, which implies, by continuity of ϕ^{-1}, that $\boldsymbol{\pi}^\star(\boldsymbol{x})$ is an accumulation point of $(\boldsymbol{\pi}_n(\boldsymbol{x}))_{n \in \mathbb{N}}$. Given the definition of $\boldsymbol{\pi}_n(\boldsymbol{x})$, it is clear by induction that for any $n \in \mathbb{N}$, $h(\boldsymbol{x}, \boldsymbol{\pi}_n(\boldsymbol{x})) \leqslant \min_{m \leqslant n} h(\boldsymbol{x}, \boldsymbol{\pi}^\star_m)$, and thus $h(\boldsymbol{x}, \boldsymbol{\pi}^\star(\boldsymbol{x})) \leqslant \min_{n \in \mathbb{N}} h(\boldsymbol{x}, \boldsymbol{\pi}_n)$. By density of $(\mathbf{a}^n)_{n \in \mathbb{N}}$ and by continuity of $h(\boldsymbol{x}, \cdot)$, this implies that $h(\boldsymbol{x}, \boldsymbol{\pi}^\star(\boldsymbol{x})) = \min_{\mathbf{a} \in A^N} h(\boldsymbol{x}, \mathbf{a})$ for all $\boldsymbol{x} \in \mathcal{X}^N$, i.e. $\boldsymbol{\pi}^\star(\boldsymbol{x}) \in \operatorname{argmin}_{\mathbf{a} \in A^N} h(\boldsymbol{x}, \mathbf{a})$. We conclude that $\boldsymbol{\pi}^\star$ is thus a measurable selection of $\operatorname{argmin}_{\mathbf{a} \in A^N} h(., \mathbf{a})$. □

By Lemma 3.4, there exists a randomized feedback policy $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon, N}$ s.t.

$$\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_{_N}[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_{_N}[\boldsymbol{x}, \boldsymbol{\pi}^{\mathfrak{a}_\epsilon, N}]\big) \quad = \quad \inf_{\mathbf{a} \in A^N} \mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_{_N}[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_{_N}[\boldsymbol{x}, \mathbf{a}]\big),$$

and the r.h.s. of the above equality is bounded by $2M_N$ from Lemma 3.2. Together with Proposition 3.1, this proves Theorem 2.2.

3.4 Proof of Theorem 2.3

Given a randomized feedback policy $\boldsymbol{\pi}_r \in L^0(\mathcal{X}^N \times [0,1]^N; A^N))$, the set of measurable functions from $\mathcal{X}^N \times [0,1]^N)$ into A^N, the associated feedback control is the unique control $\boldsymbol{\alpha}^\pi$ given by $\boldsymbol{\alpha}_t^{\boldsymbol{\pi}} = \boldsymbol{\pi}_r(\mathbf{X}_t, \boldsymbol{U}_t)$, $t \in \mathbb{N}$, where $\{\boldsymbol{U}_t = (U_t^i)_{i\in[\![1,N]\!]}, t \in \mathbb{N}\}$ is a family of mutually i.i.d. uniform random variables on $[0,1]$, independent of $\mathcal{G}$, $\boldsymbol{\varepsilon}$. By misuse of notation, we denote $V_N^{\boldsymbol{\pi}} = V_N^{\boldsymbol{\alpha}^{\boldsymbol{\pi}}}$. For $\boldsymbol{\pi}_r \in L^0(\mathcal{X}^N \times [0,1]^N; A^N)$, we introduce the operator $\mathcal{T}_N^{\boldsymbol{\pi}_r}$ on $L_m^\infty(\mathcal{X}^N)$, defined by

$$\mathcal{T}_N^{\boldsymbol{\pi}_r} W(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)) + \beta W(\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0), \boldsymbol{\varepsilon}_1)], \quad \forall \boldsymbol{x} \in \mathcal{X}^N,$$

where $\boldsymbol{U}_0 = (U_0^i)_{i\in[\![1,N]\!]}$ is a family of i.i.d. $\sim \mathcal{U}([0,1])$, independent of $\mathcal{G}$, $\boldsymbol{\varepsilon}$.

We adapt Proposition 3.1 to the case of randomized feedback policies.

Proposition 3.2 *Let $\mathfrak{a}_\epsilon$ be an ε-optimal randomized feedback policy for the CMKV-MDP, and consider any feedback policy $\boldsymbol{\pi}_r$ for the N-agent MDP. Then, the feedback control $\boldsymbol{\alpha}^{\boldsymbol{\pi}_r}$ is*

$$\mathcal{O}(\epsilon + \sup_{\boldsymbol{x}\in\mathcal{X}^N} \mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)\Big]^\gamma + M_N^\gamma)\text{-optimal for } V_N(\boldsymbol{x}_0),$$

namely

$$V_N(\boldsymbol{x}_0) - C\Big(\epsilon + \sup_{\boldsymbol{x}\in\mathcal{X}^N} \mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)\Big]^\gamma + M_N^\gamma\Big) \;\leqslant\; V_N^{\boldsymbol{\pi}}(\boldsymbol{x}_0).$$

Here $(\xi_{\boldsymbol{x}}, U) \sim \mu_N[\boldsymbol{x}] \otimes \mathcal{U}([0,1])$, and $\boldsymbol{U}_0 = (U_0^i)_{i\in[\![1,N]\!]}$ is a family of i.i.d. $\sim \mathcal{U}([0,1])$, independent of $\boldsymbol{\varepsilon}$.

Proof. Fix $\boldsymbol{x} \in \mathcal{X}^N$, and let $\mathbf{a} = \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)$ be the random variable valued in A^N. By definition, we have $\mathcal{T}_N^{\boldsymbol{\pi}_r} V_N(\boldsymbol{x}) = \mathbb{E}\big[\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x})\big]$. By Lemma 3.3, we have

$$\mathbb{T}_N^{\mathbf{a}} V_N(\boldsymbol{x}) \;\geqslant\; V_N(\boldsymbol{x}) - \epsilon - C\big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)^\gamma + M_N^\gamma\big].$$

Taking the expectation, and by Jensen's inequality, we then get

$$\mathcal{T}_N^{\boldsymbol{\pi}_r} V_N(\boldsymbol{x}) \;\geqslant\; V_N(\boldsymbol{x}) - \epsilon - C\Big(\sup_{\boldsymbol{x}\in\mathcal{X}^N} \mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)\Big]^\gamma + M_N^\gamma\Big),$$

and we conclude by the verification result in Lemma B.5. □

Compared to Proposition 3.1, Proposition 3.2 means that with a randomized feedback policy $\boldsymbol{\pi}_r$, one can obtain a "good" performance whenever it produces empirical state-action distributions that are close the theoretical state-action distribution generated by $\mathfrak{a}_\epsilon$ *on average*, i.e., that makes the quantity

$$\mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)\Big]$$

as small as possible. More precisely, if we can design a randomized policy $\boldsymbol{\pi}_r$ such that

$$\mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}}, \mathfrak{a}_\epsilon(\mu_N[\boldsymbol{x}], \xi_{\boldsymbol{x}}, U)), \mu_N[\boldsymbol{x}, \boldsymbol{\pi}_r(\boldsymbol{x}, \boldsymbol{U}_0)]\big)\Big] \;\leqslant\; CM_N,$$

then by Proposition 3.2, this will prove the statement of Theorem 2.3. The next result shows how it can be achieved.

Lemma 3.5 *Let $\mathfrak{a} : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to A$ be any (if it exists) randomized feedback policy for the CMKV-MDP such that*

$$\mathbb{E}[d_A(\mathfrak{a}(\mu,x,U),\mathfrak{a}(\mu,x',U))] \leqslant Kd(x,x'), \quad \forall \mu \in \mathcal{P}(\mathcal{X}), x, x' \in \mathcal{X}, \tag{3.9}$$

(here $U \sim \mathcal{U}[0,1]$)) for some positive constant K. Consider the randomized feedback policy for the N-agent MDP defined by

$$\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{x},\boldsymbol{u}) = \Big(\mathfrak{a}(\mu_{_N}[\boldsymbol{x}],x^i,u^i)\Big)_{i\in[\![1,N]\!]}, \quad \boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]} \in \mathcal{X}^N,\ \boldsymbol{u} = (u^i)_{i\in[\![1,N]\!]} \in [0,1]^N.$$

Then,

$$\mathbb{E}\Big[\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathfrak{a}(\mu_{_N}[\boldsymbol{x}],\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{x},\boldsymbol{U}_0)]\big)\Big] \quad \leqslant \quad (2+K)M_N,$$

where $(\xi_{\boldsymbol{x}},U) \sim \mu_{_N}[\boldsymbol{x}] \otimes \mathcal{U}([0,1])$.

Proof. Fix $\boldsymbol{x} \in \mathcal{X}^N$, and set $\mathrm{a}_{\boldsymbol{x}}(x,u) = \mathfrak{a}(\mu_{_N}[\boldsymbol{x}],x,u)$ for $(x,u) \in \mathcal{X}\times[0,1]$. Let us consider a family $\boldsymbol{\xi} = (\xi^i)_{i\in[\![1,N]\!]}$ of N i.i.d. random variables such that $\xi^i \sim \mu_{_N}[\boldsymbol{x}]$, and independent of $\boldsymbol{U}_0$. Let us consider $\sigma^{\boldsymbol{\xi},\boldsymbol{x}}$, the optimal permutation defined in Definition 3.1 between $\boldsymbol{\xi}$ and $\boldsymbol{x}$. We have

$$\begin{aligned}
&\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathfrak{a}(\mu_{_N}[\boldsymbol{x}],\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{x},\boldsymbol{U}_0)]\big)\\
=\ &\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}_{\boldsymbol{x}}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N \delta_{x^i,\mathrm{a}_{\boldsymbol{x}}(x^i,U_0^i)}\big)\\
\leqslant\ &\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}_{\boldsymbol{x}}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N \delta_{\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},\mathrm{a}_{\boldsymbol{x}}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)}\big) + \mathcal{W}_{\mathbf{d}}\big(\frac{1}{N}\sum_{i=1}^N \delta_{\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},\mathrm{a}_{\boldsymbol{x}}(\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}},U_0^i)},\frac{1}{N}\sum_{i=1}^N \delta_{x^i,\mathrm{a}_{\boldsymbol{x}}(x^i,U_0^i)}\big)\\
\leqslant\ &\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathrm{a}_{\boldsymbol{x}}(\xi_{\boldsymbol{x}},U)),\frac{1}{N}\sum_{i=1}^N \delta_{\xi^i,\mathrm{a}_{\boldsymbol{x}}(\xi^i,U_0^{(\sigma^{\boldsymbol{\xi},\boldsymbol{x}})_i^{-1}})}\big) + \mathbf{d}_N\big((\boldsymbol{\xi}^{\sigma^{\boldsymbol{\xi},\boldsymbol{x}}},\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{\xi}^{\sigma^{\boldsymbol{\xi},\boldsymbol{x}}},\boldsymbol{U}_0)),(\boldsymbol{x},\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{x},\boldsymbol{U}_0))\big),
\end{aligned}$$

where we set $\boldsymbol{\xi}^{\sigma^{\boldsymbol{\xi},\boldsymbol{x}}} = (\xi^{\sigma_i^{\boldsymbol{\xi},\boldsymbol{x}}})_{i\in[\![1,N]\!]}$, and use (2.2) in the last inequality. Taking the expectation, we then obtain under condition (3.9)

$$\begin{aligned}
&\mathcal{W}_{\mathbf{d}}\big(\mathcal{L}(\xi_{\boldsymbol{x}},\mathfrak{a}(\mu_{_N}[\boldsymbol{x}],\xi_{\boldsymbol{x}},U)),\mu_{_N}[\boldsymbol{x},\boldsymbol{\pi}_r^{\mathfrak{a},N}(\boldsymbol{x},\boldsymbol{U}_0)]\big)\\
\leqslant\ &M_N + (1+K)\mathbb{E}[\boldsymbol{d}_N(\boldsymbol{\xi}^{\sigma^{\boldsymbol{\xi},\boldsymbol{x}}},\boldsymbol{x})]\\
=\ &M_N + (1+K)\mathbb{E}\big[\mathcal{W}_d\big(\mu_{_N}[\boldsymbol{\xi}],\mu_{_N}[\boldsymbol{x}]\big)\big] \ \leqslant\ (2+K)M_N,
\end{aligned}$$

where we use (3.1) in the last equality. This concludes the proof. □

We now apply Lemma 3.5 with an ε-optimal randomized feedback policy $\mathfrak{a}_\epsilon$ for the CMKV-MDP, and combined with Proposition 3.2, this proves the required result in Theorem 2.3.

A Existence of optimal randomized control for CMKV-MDP

Recall from Proposition 4.1 in [16] that the Bellman operator $\mathcal{T}$ of the CMKV-MDP is written in the lifted form as

$$[\mathcal{T}W](\mu) = \sup_{\boldsymbol{a}\in\mathbf{A}}\Big\{\tilde{f}(\mu,\boldsymbol{a}) + \beta\mathbb{E}\big[W\big(\tilde{F}(\mu,\boldsymbol{a},\varepsilon_1^0)\big)\big]\Big\}, \quad \mu \in \mathcal{P}(\mathcal{X}), \tag{A.1}$$

for $W \in L^\infty_m(\mathcal{P}(\mathcal{X}))$, where $\mathbf{A} = \mathcal{P}(\mathcal{X} \times A)$, $\tilde{F}$ is the measurable function on $\mathcal{P}(\mathcal{X}) \times \mathbf{A} \times E^0 \to \mathcal{P}(\mathcal{X})$ defined by

$$\tilde{F}(\mu, \boldsymbol{a}, e^0) = F(\cdot, \cdot, \mathbf{p}(\mu, \boldsymbol{a}), \cdot, e^0) \star \big(\mathbf{p}(\mu, \boldsymbol{a}) \otimes \mathcal{L}(\varepsilon_1)\big),$$

and $\tilde{f}$ is the measurable function on $\mathcal{P}(\mathcal{X}) \times \mathbf{A}$ defined by

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

Here $\star$ is the pushforward measure notation, $\mathbf{p}$ is a measurable coupling projection from $\mathcal{P}(\mathcal{X}) \times \mathbf{A}$ into $\mathbf{A}$: $\mathbf{p}(\mu, \boldsymbol{a}) = \mathbf{p}(\mu, \mathbf{p}(\mu, \boldsymbol{a}))$, satisfying $\mathrm{pr}_1 \star \mathbf{p}(\mu, \boldsymbol{a}) = \mu$, and $\mathbf{p}(\mu, \boldsymbol{a}) = \boldsymbol{a}$ if $\mathrm{pr}_1 \star \boldsymbol{a} = \mu$ (where pr_1 is the projection function on the first coordinate). Since $\tilde{f}$ and $\tilde{F}$ depend upon $\boldsymbol{a}$ only through $\mathbf{p}(\mu, \boldsymbol{a})$, it is clear that the supremum in (A.1), for each $\mu \in \mathcal{P}(\mathcal{X})$, can be taken actually over the the subset $\Gamma_\mu := \{\boldsymbol{a} : (\mu, \boldsymbol{a}) \in \Gamma\} \subset \mathbf{A}$, where $\Gamma := \{(\mu, \boldsymbol{a}) \in \mathcal{P}(\mathcal{X}) \times \mathbf{A} : \mathrm{pr}_1 \star \boldsymbol{a} = \mu\}$ is closed in $\mathcal{P}(\mathcal{X}) \times \mathbf{A}$ from the continuity of $\boldsymbol{a} \mapsto \mathrm{pr}_1 \star \boldsymbol{a}$. Moreover, since V is continuous (see (3.2)), it is straightforward to prove that

$$\begin{aligned}(\mu, \boldsymbol{a}) \in \Gamma \quad \mapsto \quad & \tilde{f}(\mu, \boldsymbol{a}) + \beta \mathbb{E}\Big[V\big(\tilde{F}(\mu, \boldsymbol{a}, \varepsilon_1^0)\big)\Big] \\ & = \int_{\mathcal{X} \times A} f(x, a, \boldsymbol{a}) \boldsymbol{a}(\mathrm{d}x, \mathrm{d}a) + \beta \mathbb{E}\Big[V\big(F(\cdot, \cdot, \mu, \cdot, e^0) \star \big(\boldsymbol{a} \otimes \mathcal{L}(\varepsilon_1)\big)\big)\Big]\end{aligned}$$

is continuous and thus upper continuous on Γ. Therefore, by [2], Proposition 7.33, there exists a measurable function $\phi : \mathcal{P}(\mathcal{X}) \to \mathbf{A}$ whose graph is included in Γ and such that

$$\begin{aligned}\tilde{f}(\mu, \phi(\mu)) + \beta \mathbb{E}\big[V\big(\tilde{F}(\mu, \phi(\mu), \varepsilon_1^0)\big)\big] &= \sup_{\boldsymbol{a} \in \Gamma_\mu} \Big\{\tilde{f}(\mu, \boldsymbol{a}) + \beta \mathbb{E}\big[V\big(\tilde{F}(\mu, \boldsymbol{a}, \varepsilon_1^0)\big)\big]\Big\}, \\ &= [\mathcal{T}V](\mu) = V(\mu), \quad \forall \mu \in \mathcal{P}(\mathcal{X}), \qquad \text{(A.2)}\end{aligned}$$

where the last equality follows from the fixed point equation of V. By the universal disintegration theorem (see [13], Corollary 1.26), there exists $\kappa : \mathcal{X} \times \mathcal{P}(\mathcal{X} \times A) \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(A)$ such that for all $\boldsymbol{a} \in \mathcal{P}(\mathcal{X} \times A)$, $\mu \in \mathcal{P}(\mathcal{X})$ with $\mathrm{pr}_1 \star \boldsymbol{a} = \mu$, we have $\boldsymbol{a} = \mu \hat{\otimes} \kappa(\cdot, \boldsymbol{a}, \mu)$ (where $\hat{\otimes}$ denotes the probability-kernel product). Furthermore, by Blackwell-Dubins Lemma, there exists a measurable function $\rho : \mathcal{P}(A) \times [0, 1] \to A$ such that for all $\pi \in \mathcal{P}(A)$, if U denotes a uniform random variable, then $\rho(\pi, U) \sim \pi$. We can then define the randomized feedback policy

$$\mathfrak{a}_0(\mu, x, u) \quad = \quad \rho(\kappa(x, \phi(\mu), \mu), u),$$

which satisfies by construction $\mathcal{L}(\xi, \mathfrak{a}_0(\mu, \xi, U)) = \phi(\mu)$ for $(\xi, U) \sim \mu \otimes \mathcal{U}([0, 1])$ so that

$$\begin{aligned}\tilde{f}(\mu, \phi(\mu)) \quad &= \quad \mathbb{E}\Big[f\big(\xi, \mathfrak{a}_0(\mu, \xi, U), \mathcal{L}(\xi, \mathfrak{a}_0(\mu, \xi, U))\big)\Big] \\ \tilde{F}(\mu, \phi(\mu), \varepsilon_1^0) \quad &= \quad \mathbb{P}^0_{F(\xi, \mathfrak{a}_0(\mu, \xi, U), \mathcal{L}(\xi, \mathfrak{a}_0(\mu, \xi, U)), \varepsilon_1, \varepsilon_1^0)}.\end{aligned}$$

Recalling notation in (2.4), and by (A.2), this shows that

$$\mathbb{T}^{\mathfrak{a}_0(\mu, \cdot)} V(\mu) = V(\mu).$$

According to the verification result (Proposition 4.3 in [16]), this ensures that that the randomized feedback control $\alpha^0 \in \mathcal{A}$ defined by

$$\alpha_t^0 = \mathfrak{a}_0(\mathbb{P}^0_{X_t}, X_t, U_t), \quad t \in \mathbb{N},$$

where $(U_t)_{t \in \mathbb{N}}$ is an i.i.d. sequence of random variables, $U_t \sim \mathcal{U}([0, 1])$, independent of $\xi_0 \sim \mu_0$, and $\boldsymbol{\varepsilon}$, is an optimal control for $V(\mu_0)$.

B Bellman equation for the N-agent MDP

In this section, we study and rigorously state properties on the Bellman equation for the N-agent problem, viewed as a MDP with state space $\mathcal{X}^N$, action space A^N, noise sequence $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_t)_{t\in\mathbb{N}^\star}$ with $\boldsymbol{\varepsilon}_t := ((\varepsilon_t^i)_{i\in[\![1,N]\!]}, \varepsilon_t^0)$ valued in $E^N \times E^0$, state transition function

$$\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, \mathbf{e}) \;\; := \;\; \Big(F(x^i, a^i, \mu_{\scriptscriptstyle N}[\boldsymbol{x}, \mathbf{a}], e^i, e^0)\Big)_{i\in[\![1,N]\!]}, \quad \mathbf{e} = ((e^i)_{i\in[\![1,N]\!]}, e^0) \in E^N \times E^0,$$

and reward function

$$\boldsymbol{f}(\boldsymbol{x}, \mathbf{a}) \;\; = \;\; \frac{1}{N}\sum_{i=1}^N f\big(x^i, a^i, \mu_{\scriptscriptstyle N}[\boldsymbol{x}, \mathbf{a}]\big), \quad \boldsymbol{x} = (x^i)_{i\in[\![1,N]\!]},\ \mathbf{a} = (a^i)_{i\in[\![1,N]\!]}.$$

With respect to standard framework of MDP, we pay a careful attention when dealing with possibly continuous state/action spaces $(\mathcal{X}, A)$, and optimizing in general over open-loop controls.

Let us consider the set $\boldsymbol{\mathcal{V}}$ of sequences $\boldsymbol{\nu} = (\boldsymbol{\nu}_t)_{t\in\mathbb{N}}$ with $\boldsymbol{\nu}_0$ a measurable function from $([0,1]^N)^{\mathbb{N}}$ into A^N, and $\boldsymbol{\nu}_t$ a measurable function from $([0,1]^N)^{\mathbb{N}} \times (E^N \times E^0)^t$ into A^N for $t \in \mathbb{N}^\star$. For each $\boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$, we can associate a control process $\boldsymbol{\alpha}^{\boldsymbol{\nu}} \in \boldsymbol{\mathcal{A}}$ given by

$$\boldsymbol{\alpha}_t^{\boldsymbol{\nu}} \;\; := \;\; \boldsymbol{\nu}_t(\boldsymbol{U}, (\boldsymbol{\varepsilon}_s)_{s\in[\![1,t]\!]}), \quad t \in \mathbb{N},$$

(with the convention that $\boldsymbol{\alpha}_0^{\boldsymbol{\nu}} = \boldsymbol{\nu}_0(\boldsymbol{U})$ when $t = 0$), where $\boldsymbol{U} = (U_t^i)_{i\in[\![1,N]\!],t\in\mathbb{N}}$ is a family of mutually i.i.d. uniform random variables on $[0,1]$, independent of $\boldsymbol{\varepsilon}$, and conversely any control $\boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}$ can be represented as $\boldsymbol{\alpha}^{\boldsymbol{\nu}}$ for some $\boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$. We call $\boldsymbol{\mathcal{V}}$ the set of randomized open-loop policies. By misuse of notation, we write $V_N^{\boldsymbol{\nu}} = V_N^{\boldsymbol{\alpha}^{\boldsymbol{\nu}}}$.

Let us denote by $L^\infty(\mathcal{X}^N)$ the set of bounded real-valued functions on $\mathcal{X}^N$, and by $L_m^\infty(\mathcal{X}^N)$ the subset of measurable functions in $L^\infty(\mathcal{X}^N)$. We then introduce the Bellman "operator" $\mathcal{T}_N : L_m^\infty(\mathcal{X}^N) \to L^\infty(\mathcal{X}^N)$ defined for any $W \in L_m^\infty(\mathcal{X}^N)$ by:

$$[\mathcal{T}_N W](\boldsymbol{x}) := \sup_{\mathbf{a}\in A^N} \mathbb{T}_N^{\mathbf{a}} W(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}^N.$$

where

$$\mathbb{T}_N^{\mathbf{a}} W(\boldsymbol{x}) := \; \boldsymbol{f}(\boldsymbol{x}, \mathbf{a}) + \beta\mathbb{E}\big[W\big(\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, \boldsymbol{\varepsilon}_1)\big)\big], \quad \boldsymbol{x} \in \mathcal{X}^N,\ \mathbf{a} \in A^N.$$

Notice that the sup can a priori lead to a non measurable function $\mathcal{T}_N W$. Because of this, $\mathcal{T}_N$ is not an operator on $L_m^\infty(\mathcal{X}^N)$ in the strict sense. To see $\mathcal{T}_N$ as an operator, we have to find a subset in $L_m^\infty(\mathcal{X}^N)$ that is preserved by $\mathcal{T}_N$. The next result introduces such subset.

Lemma B.1 *Let $\mathcal{M}$ be the set in $L_m^\infty(\mathcal{X}^N)$ defined by*

$$\mathcal{M} := \; \Big\{W \in L_m^\infty(\mathcal{X}^N) : \big|W(\boldsymbol{x}) - W(\boldsymbol{x}')\big| \leqslant 2K_f \sum_{t=0}^\infty \beta^t \min\big[(2K_F)^t \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\big],\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^N\Big\}. \tag{B.1}$$

Then $\mathcal{M}$ is a complete metric space under the $\|\cdot\|$ norm, and $\mathbb{T}_N^{\mathbf{a}}$, for all $\mathbf{a} \in A^N$, and $\mathcal{T}_N$, preserve $\mathcal{M}$: $\mathbb{T}_N^{\mathbf{a}}\mathcal{M} \subset \mathcal{M}$, $\mathcal{T}_N\mathcal{M} \subset \mathcal{M}$.

Proof. It is clear that $\mathcal{M}$ is closed in $L^\infty_m(\mathcal{X}^N)$, and is therefore a complete metric space for $\|\cdot\|$. Let $W \in \mathcal{M}$. Fix $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^N$, and $\mathbf{a} \in A^N$. Let us start with two preliminary estimations: under $(\mathbf{Hf_{lip}})$, and recalling (2.2), we clearly have

$$|\boldsymbol{f}(\boldsymbol{x}, \mathbf{a}) - \boldsymbol{f}(\boldsymbol{x}', \mathbf{a})| \leqslant 2K_f d_N(\boldsymbol{x}, \boldsymbol{x}'). \tag{B.2}$$

Similarly, under $(\mathbf{HF_{lip}})$, for $e^0 \in E^0$, we have

$$\mathbb{E}[\boldsymbol{d}_N(\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e^0), \boldsymbol{F}(\boldsymbol{x}', \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e^0))] \leqslant 2K_F \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'). \tag{B.3}$$

Thus, denoting by $\mathbf{X}_1 = \boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e^0)$ and $\mathbf{X}_1' = \boldsymbol{F}(\boldsymbol{x}', \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e^0)$, we have, by Jensen's inequality and then (B.3),

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^\infty \beta^t \min\left[(2K_F)^t \boldsymbol{d}_N(\mathbf{X}_1, \mathbf{X}_1'), \Delta_{\mathcal{X}}\right]\right] &\leqslant \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^t \mathbb{E}[\boldsymbol{d}_N(\mathbf{X}_1, \mathbf{X}_1')], \Delta_{\mathcal{X}}\right] \\ &\leqslant \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^{t+1} \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\right]. \end{aligned} \tag{B.4}$$

The definition of $\mathbb{T}^{\mathbf{a}}_N$ combined with (B.2), the fact that $W \in \mathcal{M}$, and (B.4), implies that

$$\begin{aligned}|\mathbb{T}^{\mathbf{a}}_N W(x) - \mathbb{T}^{\mathbf{a}}_N W(x')| &\leqslant 2K_f \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}') + \beta 2K_f \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^{t+1} \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\right] \\ &\leqslant 2K_f \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^t \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\right],\end{aligned}$$

which shows that $\mathbb{T}^{\mathbf{a}}_N W \in \mathcal{M}$, i.e. $\mathbb{T}^{\mathbf{a}}_N$ preserves $\mathcal{M}$. Furthermore, we have

$$\begin{aligned}|\mathcal{T}_N W(\boldsymbol{x}) - \mathcal{T}_N W(\boldsymbol{x}')| &\leqslant \sup_{\mathbf{a}\in A^N} |\mathbb{T}^{\mathbf{a}}_N W(\boldsymbol{x}) - \mathbb{T}^{\mathbf{a}}_N W(\boldsymbol{x}')| \\ &\leqslant 2K_f \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^t \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\right],\end{aligned}$$

which also shows that $\mathcal{T}_N W \in \mathcal{M}$. □

Lemma B.1 implies that by restricting $\mathcal{T}_N$ and $\mathbb{T}^{\mathbf{a}}_N$ to $\mathcal{M}$, we can see $\mathcal{T}_N$ and $\mathbb{T}^{\mathbf{a}}_N$ as operators on $\mathcal{M}$, that is, $\mathcal{T}_N : \mathcal{M} \to \mathcal{M}$ and $\mathbb{T}^{\mathbf{a}}_N : \mathcal{M} \to \mathcal{M}$. However, the property defining the functions in $\mathcal{M}$ (see (B.1)) is not very natural and practical. The following result provides a more convenient property satisfied by all functions in $\mathcal{M}$.

Lemma B.2 *There exists $K_\star \in \mathbb{R}$ such that any function $W \in \mathcal{M}$ is γ-Hölder with constant factor $K_\star$, i.e.*

$$\left|W(\boldsymbol{x}) - W(\boldsymbol{x}')\right| \leqslant K_\star \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}')^\gamma, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^N.$$

Proof. We have

$$|W(\boldsymbol{x}) - W(\boldsymbol{x}')| \leqslant 2K_f \sum_{t=0}^\infty \beta^t \min\left[(2K_F)^t \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}'), \Delta_{\mathcal{X}}\right] =: 2K_f S(\boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}')).$$

where $S(m) = \sum_{t=0}^{\infty} \beta^t \min[(2K_F)^t m, \Delta_{\mathcal{X}}]$. If $2\beta K_F < 1$, we clearly have

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

and so W is 1-Hölder. Let us now study the case $2\beta K_F > 1$. In this case, in particular, $2K_F > 1$ since $\beta \in (0,1)$, thus $t \mapsto (2K_F)^t$ is nondecreasing, and so

$$\begin{aligned} S(m) &\leqslant \sum_{t=0}^{\infty} \int_t^{t+1} \beta^t \min\big[(2K_F)^s m, \Delta_{\mathcal{X}}\big] ds \\ &\leqslant \frac{1}{\beta} \sum_{t=0}^{\infty} \int_t^{t+1} \beta^s \min\big[(2K_F)^s m, \Delta_{\mathcal{X}}\big] ds \;=\; \frac{1}{\beta} \int_0^{\infty} e^{s \ln \beta} \min\big[m e^{s \ln(2K_F)}, \Delta_{\mathcal{X}}\big] ds. \end{aligned}$$

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

$$\begin{aligned} \int_0^{\infty} e^{s \ln \beta} \min\big[m e^{s \ln(2K_F)}, \Delta_{\mathcal{X}}\big] ds &\leqslant m \int_0^{t^\star} e^{s \ln(2K_F \beta)} ds + \Delta_{\mathcal{X}} \int_{t^\star}^{\infty} e^{s \ln(\beta)} ds \\ &= \frac{m}{\ln(2K_F\beta)} \Big[e^{t_\star \ln(2K_F\beta)} - 1 \Big] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} e^{\ln(\beta) t^\star} \\ &= \frac{m}{\ln(2K_F\beta)} \Big[\Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(2K_F\beta)}{\ln(2K_F)}} - 1 \Big] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} \left(\frac{\Delta_{\mathcal{X}}}{m}\right)^{\frac{\ln(\beta)}{\ln(2K_F)}} \\ &= \Delta_{\mathcal{X}} \Big(\frac{1}{\ln(2K_F\beta)} - \frac{1}{\ln \beta} \Big) \Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(\beta)}{\ln(2K_F)}} - \frac{m}{\ln(2K_F\beta)} \\ &\leqslant C m^{\min\left[1, \frac{|\ln \beta|}{\ln(2K_F)}\right]} = C m^{\gamma}, \end{aligned}$$

for some positive constant C depending on K_F, β and $\Delta_{\mathcal{X}}$. This implies that W is γ-Hölder with a constant factor $K_\star$ that is clearly independent of $W \in \mathcal{S}$. This concludes the proof. □

The consequence of Lemmas B.1 and B.2 is that the set $\mathcal{M} \subset L_m^\infty(\mathcal{X})$ is a closed set, preserved by $\mathcal{T}_N$ and contains only functions that are γ-Hölder with factor $K_\star$. We are now able to get the existence of a unique fixed point to the Bellman operator $\mathcal{T}_N$.

Proposition B.1 *(i) The operator $\mathcal{T}_N$ is monotone increasing: for $W_1, W_2 \in L_m^\infty(\mathcal{X}^N)$, if $W_1 \leqslant W_2$, then $\mathcal{T}_N W_1 \leqslant \mathcal{T}_N W_2$. (ii) Furthermore, it is contracting on $L_m^\infty(\mathcal{X}^N)$ with Lipschitz factor β, and admits a unique fixed point in $L_m^\infty(\mathcal{X}^N)$, denoted by $V_N^\star$, hence solution to:*

$$V_N^\star = \mathcal{T}_N V_N^\star.$$

Moreover, $V_N^\star \in \mathcal{M}$, and thus $V_N^\star$ is γ-Hölder with constant factor $K_\star$.

Proof. (i) The monotonicity of $\mathcal{T}_N$ is clear. (ii) The β-contraction property of $\mathcal{T}_N$ is obtained by standard arguments, which implies the uniqueness of a fixed point (but not the existence). Let us prove the existence of a fixed point. As $\mathcal{M}$ is preserved by $\mathcal{T}_N$, and is closed for $\|\cdot\|$, and therefore complete (as a closed subset of the complete space $L_m^\infty(\mathcal{X}^N)$), by the Banach fixed point theorem, $\mathcal{T}_N$ admits a unique fixed point $V_N^\star$ in $\mathcal{M}$. By Lemma B.2, this implies that $V_N^\star$ is γ-Hölder with constant factor $K_\star$, and concludes the proof. □

Remark B.1 Notice that the above arguments would not work if we considered, instead of $\mathcal{M}$, directly the set of γ-Hölder continuous functions. Indeed, while it is true that such set is stabilized by $\mathcal{T}_N$ (it essentially follows from (B.2) and (B.3)), the set of γ-Hölder continuous functions is not closed in $L^\infty_m(\mathcal{X}^N)$ (and thus not a complete metric space): there might indeed exist a converging sequence of γ-Hölder continuous functions with multiplicative factors (in the Hölder property) tending toward infinity, such that the limit function is not γ-Hölder anymore. □

As a consequence of Proposition B.1, we can show the following relation between the value function V_N of the N-agent MDP, and the fixed point $V_N^\star$ of the Bellman operator $\mathcal{T}_N$.

Lemma B.3 *For all $\boldsymbol{x} \in \mathcal{X}^N$, we have $V_N(\boldsymbol{x}) \leqslant V_N^\star(\boldsymbol{x})$.*

Proof. For any $\boldsymbol{x} \in \mathcal{X}^N$, $\boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$, we have

$$\begin{aligned}\mathbb{E}\Big[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{U})) + \beta V_N^\star\big(\boldsymbol{F}(x,\boldsymbol{\nu}_0(\boldsymbol{U}),\boldsymbol{\varepsilon}_1)\big)\Big] &= \mathbb{E}\Big[\Big\{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\nu}_0(u)) + \beta\mathbb{E}[V_N^\star(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{u}),\boldsymbol{\varepsilon}_1))]\Big\}_{\boldsymbol{u}:=\boldsymbol{U}}\Big] \\ &= \mathbb{E}\Big[\mathbb{T}^{\boldsymbol{\nu}_0(\boldsymbol{U})}V_N^\star(\boldsymbol{x})\Big] \leqslant \mathcal{T}V_N^\star(\boldsymbol{x}) = V_N^\star(\boldsymbol{x}). \qquad \text{(B.5)}\end{aligned}$$

For any $(\boldsymbol{u},\mathbf{e}) \in ([0,1]^N)^{\mathbb{N}} \times (E^N \times E^0)$, and for any $\boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$, we define $\vec{\boldsymbol{\nu}}^{\boldsymbol{u},\mathbf{e}} \in \boldsymbol{\mathcal{V}}$ by

$$\vec{\boldsymbol{\nu}}_t^{\boldsymbol{u},\mathbf{e}}(\boldsymbol{u}',(\mathbf{e}'_s)_{s\in[\![1,t]\!]}) := \boldsymbol{\nu}_{t+1}(\boldsymbol{u},\mathbf{e},(\mathbf{e}'_s)_{s\in[\![1,t]\!]}), \quad (\boldsymbol{u}',(\mathbf{e}'_s)_{s\in[\![1,t]\!]}) \in ([0,1]^N)^{\mathbb{N}} \times (E^N \times E^0)^t, t \in \mathbb{N}.$$

Standard Markov arguments imply the following flow property for randomized open-loop policies:

$$V_N^{\boldsymbol{\nu}}(\boldsymbol{x}) \;=\; \mathbb{E}\Big[\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{U})) + \beta V_N^{\vec{\boldsymbol{\nu}}^{\boldsymbol{U},\boldsymbol{\varepsilon}_1}}(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{U}),\boldsymbol{\varepsilon}))\Big].$$

Together with (B.5), we then get

$$\begin{aligned}V_N^\star(\boldsymbol{x}) - V_N^{\boldsymbol{\nu}}(\boldsymbol{x}) &\geqslant \beta\mathbb{E}\Big[V_N^\star(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{U}),\boldsymbol{\varepsilon}_1) - V_N^{\vec{\boldsymbol{\nu}}^{\boldsymbol{U},\boldsymbol{\varepsilon}_1}}(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\nu}_0(\boldsymbol{U}),\boldsymbol{\varepsilon}_1))\Big] \\ &\geqslant \beta \inf_{\boldsymbol{x}\in\mathcal{X}^N,\boldsymbol{\nu}\in\boldsymbol{\mathcal{V}}}\big\{V_N^\star(\boldsymbol{x}) - V_N^{\boldsymbol{\nu}}(\boldsymbol{x})\big\}.\end{aligned}$$

Taking the infimum over $\boldsymbol{x} \in \mathcal{X}^N, \boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$ on the left hand side of the above inequality, and since $\beta < 1$, this shows that $V_N^{\boldsymbol{\nu}}(\boldsymbol{x}) \leqslant V_N^\star(\boldsymbol{x})$ for all $\boldsymbol{\nu} \in \boldsymbol{\mathcal{V}}$. We conclude that $V_N \leqslant V_N^\star$. □

We aim now to prove rigorously the equality $V_N = V_N^\star$, i.e., the value function V_N of the N-agent MDP satisfies the Bellman fixed point equation: $V_N = \mathcal{T}_N V_N$, and also to show the existence of ε-optimal (randomized) feedback control for V_N.

A feedback policy (resp. randomized feedback policy) is an element $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; A^N)$ (resp. $L^0(\mathcal{X}^N \times [0,1]^N; A^N)$), the set of measurable functions from $\mathcal{X}^N$ (resp. $\mathcal{X}^N \times [0,1]^N$) into A^N. The associated feedback control is the unique control $\boldsymbol{\alpha}^\pi$ given by $\boldsymbol{\alpha}_t^{\boldsymbol{\pi}} = \boldsymbol{\pi}(\mathbf{X}_t)$, (resp. $\boldsymbol{\pi}_r(\mathbf{X}_t,\boldsymbol{U}_t)$), $t \in \mathbb{N}$, where $\{\boldsymbol{U}_t = (U_t^i)_{i\in[\![1,N]\!]}, t \in \mathbb{N}\}$ is a family of mutually i.i.d. uniform random variables on $[0,1]$, independent of $\mathcal{G}$, $\boldsymbol{\varepsilon}$. By misuse of notation, we denote $V_N^{\boldsymbol{\pi}} = V_N^{\boldsymbol{\alpha}^{\boldsymbol{\pi}}}$. Given $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; A^N)$ (resp. $L^0(\mathcal{X}^N \times [0,1]^N; A^N)$), we introduce the operator $\mathcal{T}_N^{\boldsymbol{\pi}}$ on $L^\infty_m(\mathcal{X}^N)$, defined by

$$\mathcal{T}_N^{\boldsymbol{\pi}} W(\boldsymbol{x}) := \; \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x})) + \beta\mathbb{E}\big[W\big(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{\pi}(\boldsymbol{x}),\boldsymbol{\varepsilon}_1)\big)\big], \quad \boldsymbol{x} \in \mathcal{X}^N,$$

resp.

$$\mathcal{T}_N^{\boldsymbol{\pi}} W(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{U}_0)) + \beta W(\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{U}_0), \boldsymbol{\varepsilon}_1)], \quad \forall \boldsymbol{x} \in \mathcal{X}^N,$$

where $\boldsymbol{U}_0 = (U_0^i)_{i \in [\![1,N]\!]}$ is a family of i.i.d. $\sim \mathcal{U}([0,1])$, independent of $\mathcal{G}$, $\boldsymbol{\varepsilon}$.

We have the basic and standard properties on the operator $\mathcal{T}_N^{\boldsymbol{\pi}}$:

Lemma B.4 *Fix $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; A^N)$ (resp. $L^0(\mathcal{X}^N \times [0,1]^N; A^N)$).*

(i) The operator $\mathcal{T}_N^{\boldsymbol{\pi}}$ is β-contracting on $L_m^\infty(\mathcal{X}^N)$, and $V_N^{\boldsymbol{\pi}}$ is its unique fixed point.

(ii) Furthermore, it is monotone increasing: for $W_1, W_2 \in L^\infty(\mathcal{X}^N)$, if $W_1 \leqslant W_2$, then $\mathcal{T}_N^{\boldsymbol{\pi}} W_1 \leqslant \mathcal{T}_N^{\boldsymbol{\pi}} W_2$.

We state the standard verification type result for the N-individual MDP, by means of the Bellman operator.

Lemma B.5 (Verification result)
Fix $\epsilon \geqslant 0$, and suppose that there exists an ϵ-optimal (randomized) feedback policy $\boldsymbol{\pi}^\epsilon$ for $V_N^\star$ in the sense that

$$V_N^\star \leqslant \mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star + \epsilon.$$

Then, $\boldsymbol{\alpha}^{\pi^\epsilon} \in \boldsymbol{\mathcal{A}}$ is $\frac{\epsilon}{1-\beta}$-optimal for V_N, i.e., $V_N^{\boldsymbol{\pi}^\epsilon} \geqslant V_N - \frac{\epsilon}{1-\beta}$, and we have $V_N \geqslant V_N^\star - \frac{\epsilon}{1-\beta}$.

Proof. Since $V_N^{\boldsymbol{\pi}^\epsilon} = \mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^{\boldsymbol{\pi}^\epsilon}$, and recalling from Lemma B.3 that $V_N^\star \geqslant V_N \geqslant V_N^{\boldsymbol{\pi}^\epsilon}$, we have for all $\boldsymbol{x} \in \mathcal{X}^N$,

$$\left|(V_N^\star - V_N^{\boldsymbol{\pi}^\epsilon})(\boldsymbol{x})\right| \leqslant \left|\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star(\boldsymbol{x}) - \mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^{\boldsymbol{\pi}^\epsilon}(\boldsymbol{x})\right| + \epsilon \leqslant \beta \|V_N^\star - V_N^{\boldsymbol{\pi}^\epsilon}\| + \epsilon,$$

where we used the β-contraction property of $\mathcal{T}_N^{\boldsymbol{\pi}_\epsilon}$ in Lemma B.4. We deduce that $\|V_N^\star - V_N^{\boldsymbol{\pi}^\epsilon}\| \leqslant \frac{\epsilon}{1-\beta}$, and then, $V_N \geqslant V_N^{\boldsymbol{\pi}^\epsilon} \geqslant V_N^\star - \frac{\epsilon}{1-\beta}$, which combined with $V_N^\star \geqslant V_N$, concludes the proof. □

We finally conclude this section by showing the existence of an ε-optimal (randomized) feedback policy for N-agent MDP on $\mathcal{X}^N$, and obtain as a by-product the corresponding Bellman fixed point equation for its value function.

Proposition B.2 *For all $\epsilon > 0$, there exists a (randomized) feedback policy $\boldsymbol{\pi}^\epsilon$ that is ϵ-optimal for $V_N^\star$. Consequently, the control $\boldsymbol{\alpha}^{\boldsymbol{\pi}^\epsilon} \in \boldsymbol{\mathcal{A}}$ is $\frac{\epsilon}{1-\beta}$-optimal for V_N, and we have $V_N = V_N^\star$, which thus satisfies the Bellman fixed point equation.*

Proof. We prove the result for ϵ-optimal feedback policy (the case of ϵ-optimal randomized feedback policy is dealt with similarly). Fix $\epsilon > 0$, and given $\eta > 0$, consider a quantizing grid $\mathcal{M}^\eta = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N^\eta}\} \subset \mathcal{X}^N$, and an associated partition C_k^η, $k = 1, \ldots, N^\eta$, of $\mathcal{X}^N$, satisfying

$$C_k^\eta \subset B^\eta(\boldsymbol{x}_k) := \Big\{\boldsymbol{x} \in \mathcal{X}^N : \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}_k) \leqslant \eta\Big\}, \quad k = 1, \ldots, N_\eta.$$

For any $\boldsymbol{x}_k$, $k = 1, \ldots, N^\eta$, there exists $\mathbf{a}_k^\varepsilon \in A^N$ such that

$$V_N^\star(\boldsymbol{x}_k) \leqslant \mathbb{T}^{\mathbf{a}_k^\varepsilon} V_N^\star(\boldsymbol{x}_k) + \frac{\epsilon}{3}. \tag{B.6}$$

From the partition C_k^η, $k = 1, \ldots, N_\eta$ of $\mathcal{X}^N$, associated to $\mathcal{M}^\eta$, we construct the function $\boldsymbol{\pi}^\epsilon : \mathcal{X}^N \to A^N$ as follows: we define, for all $\boldsymbol{x} \in \mathcal{X}^N$,

$$\pi^\epsilon(\boldsymbol{x}) \quad = \quad \mathbf{a}_k^\epsilon, \quad \text{when } \boldsymbol{x} \in C_k^\eta,\ k = 1, \ldots, N^\eta.$$

Such function $\boldsymbol{\pi}^\epsilon$ is clearly measurable. Let us now check that such $\boldsymbol{\pi}^\epsilon$ yields an ϵ-optimal feedback policy for η small enough. For $\boldsymbol{x} \in \mathcal{X}^N$, we define $\boldsymbol{x}^\eta = \boldsymbol{x}_k$, when $\boldsymbol{x} \in C_k^\eta$, $k = 1, \ldots, N_\eta$. Observe that $\boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}^\eta) \leqslant \eta$. We then write for any $\boldsymbol{x} \in \mathcal{X}^N$,

$$\begin{aligned}
[\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) - V_N^\star(\boldsymbol{x}) = \ & \Big([\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) - [\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}^\eta)\Big) + \Big([\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}^\eta) - V_N^\star(\boldsymbol{x}^\eta)\Big) \\
& + \big(V_N^\star(\boldsymbol{x}^\eta) - V_N^\star(\boldsymbol{x})\big) \\
\geqslant \ & \Big([\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) - [\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}^\eta)\Big) - \frac{\epsilon}{3} - \frac{\epsilon}{3}, \qquad \text{(B.7)}
\end{aligned}$$

where we used (B.6) and the fact that $|V_N^\star(\boldsymbol{x}^\eta) - V_N^\star(\boldsymbol{x})| \leqslant \epsilon/3$ for η small enough by uniform continuity of $V_N^\star$ in Proposition B.1. Moreover, by observing that $\boldsymbol{\pi}^\epsilon(\boldsymbol{x}) = \boldsymbol{\pi}^\epsilon(\boldsymbol{x}^\eta) =: \mathbf{a}$, we have

$$\begin{aligned}
[\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) = \ & \mathbb{E}\Big[\boldsymbol{f}(\boldsymbol{x}, \mathbf{a}) + \beta V_N^\star(\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, \boldsymbol{\varepsilon}_1))\Big], \\
[\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}^\eta) = \ & \mathbb{E}\Big[\boldsymbol{f}(\boldsymbol{x}^\eta, \mathbf{a}) + \beta V_N^\star(\boldsymbol{F}(\boldsymbol{x}^\eta, \mathbf{a}, \boldsymbol{\varepsilon}_1))\Big].
\end{aligned}$$

Under ($\mathbf{HF_{lip}}$)-($\mathbf{Hf_{lip}}$), and by using the γ-Hölder property of $V_N^\star$ with constant $K_\star$ in Proposition B.1, we then get

$$\begin{aligned}
& \big|[\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) - [\mathcal{T}_N^{\boldsymbol{\pi}_\epsilon} V_N^\star](\boldsymbol{x}^\eta)\big| \\
\leqslant \ & 2K_f \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}^\eta) + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[\boldsymbol{d}_N\big(\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e), \boldsymbol{F}(x_\eta, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e)\big)^\gamma\big]_{e:=\varepsilon_1^0}\Big] \\
\leqslant \ & 2K_f \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}^\eta) + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[\boldsymbol{d}_N\big(\boldsymbol{F}(\boldsymbol{x}, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e), \boldsymbol{F}(\boldsymbol{x}^\eta, \mathbf{a}, (\varepsilon_1^i)_{i\in[\![1,N]\!]}, e)\big)\big]_{e:=\varepsilon_1^0}\Big]^\gamma \\
\leqslant \ & C \boldsymbol{d}_N(\boldsymbol{x}, \boldsymbol{x}^\eta)^\gamma \ \leqslant \ C\eta^\gamma.
\end{aligned}$$

for some constant C. Therefore, $\big|[\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}) - [\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star](\boldsymbol{x}^\eta)\big| \leqslant \epsilon/3$, and, plugging into (B.7), we obtain $\mathcal{T}_N^{\boldsymbol{\pi}^\epsilon} V_N^\star(\boldsymbol{x}) - V_N^\star(\boldsymbol{x}) \geqslant -\epsilon$, for all $\boldsymbol{x} \in \mathcal{X}^N$, which means that $\boldsymbol{\pi}^\epsilon$ is ϵ-optimal for $V_N^\star$. The rest of the assertions in the Theorem follows from the verification result in Lemma B.5. □

References

[1] N. Bäuerle. Mean Field Markov Decision Processes. *arXiv preprint arXiv:2106.08755, to appear in Applied Mathematics and Optimization*, 2021.

[2] D. Bertsekas and S. E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

[3] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[4] P. Cardaliaguet, S. Daudin, J. Jackson, and P. E. Souganidis. An algebraic convergence rate for the optimal control of Mckean-Vlasov dynamics. *arXiv:2203.14554*, 2022.

[5] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications, vol I and II.* Probability theory and stochastic modelling. Springer, 2018.

[6] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[7] A. Cecchin. Finite state N-agent and mean field control problems. *ESAIM: Control, Optimization and Calculus of Variations,*, 27(31), 2021.

[8] M. F. Djete. Extended mean field control problem: a propagation of chaos result. *arXiv preprint arXiv:2006.12996, to appear in Electronic Journal of Probability*, 2020.

[9] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[10] W. Gangbo, S. Mayorga, and A. Swiech. Finite dimensional approximations of Hamilton-Jacobi-Bellman equations in spaces of probability measures. *SIAM. J. Math. Anal*, 2:1320–1356, 2021.

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

[12] M. Germain, H. Pham, and X. Warin. Rate of convergence for particles approximation of PDEs in Wasserstein space. *arXiv:2103.00837, to appear in Journal of Applied Probability*, 2021.

[13] O. Kallenberg. *Random measures, theory and applications*, volume 1. Springer, 2017.

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

[15] M. Motte. *Mathematical models for large populations, behavorial economics, and targeted advertising*. PhD thesis, Université Paris Cité, 2021.

[16] M. Motte and H. Pham. Mean-field Markov decision processes with common noise and open-loop controls. *arXiv:1912.07883, to appear in Annals of Applied Probability*, 2019.

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

[18] A.S. Sznitman. Topics in propagation of chaos. In Springer, editor, *Ecole d'été de probabilités de Saint Flour, lecture notes in Mathematics*, volume 1464, pages 165–251. 1989.

[19] M. Thorpe. Introduction to optimal transport, 2018.