Non-Exchangeable Mean Field Markov Decision Processes with common noise : from Bellman equation to quantitative propagation of chaos

Samy Mekkaoui

GT MathsFi , CMAP 2025

Joint work with
Huyên Pham (CMAP, École Polytechnique)

22 October 2025

Contents

1. Introduction
   - Context and motivations
   - Problem formulation
2. Bellman fixed point on the space $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Lifted MDP on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Bellman fixed point on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Equivalence with the strong formulation
3. Propagation of chaos from V_N towards V
   - The N-agent problem as a MDP on the state space $\mathcal{X}^N$ and action space A^N
   - Convergence of value functions
   - Approximate optimal policies

Mean field approach to large population stochastic control

- Large number of agents N interacting dynamic agents/entities with heterogeneous interactions.
- Agents are cooperative and act following a social planner.
- When $N \to \infty$, we get an optimal control of mean-field type.
    - Symmetric agents $\to$ McKean-Vlasov equations
    - Nonsymmetric agents $\to$ New limiting systems.
- Here, we focus on
    - **Discrete time**, and finite / continuous state space
    - **Infinite Horizon**
    - Common noise
    - When $N \to \infty$: Conditional Non exchangeable Markov Decision Process (CNEMF-MDP).

$\to$ Mathematical framework of reinforcement learning (RL) with many interacting cooperative agents.

Framework and notations

- Universal filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.
- State and action spaces: $\mathcal{X}$ and A (compact and Polish) and $I = [0,1]$ encoding heterogeneity of the agents labeled by $u \in I$.
  - $\mathcal{P}(I \times \mathcal{X})$, resp $\mathcal{P}(A)$, resp $\mathcal{P}(I \times \mathcal{X} \times A)$: set of probability measures on $I \times \mathcal{X}$, resp A, resp $I \times \mathcal{X} \times A$, with Wasserstein distance.
- Discrete time **transition dynamics**
  - Idiosyncratic noises: $(\epsilon_t^u)_{u \in I}$, i.i.d valued in E.
  - Common noise: $(\epsilon_t^0)_{t \in \mathbb{N}}$ for all agents, i.i.d valued in E^0.
  - F measurable function from $I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \times E \times E^0 \to \mathcal{X}$.
- **Reward** on infinite horizon.
  - Discount factor $\beta \in [0,1)$.
  - f measurable bounded function from $I \times \mathcal{X} \times \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \to \mathbb{R}$.

The conditional McKean-Vlasov MDP problem

Conditional McKean-Vlasov Markov Decision Processes (CMKV-MDP) problem studied by Motte and Pham (see [1]):

$$V(\xi) = \inf_{\alpha \in \mathcal{A}} V^\alpha(\xi) := \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t f(X_t, \alpha_t, \mathbb{P}^0_{(X_t, \alpha_t)}) \Big], \tag{1}$$

where $\mathcal{A}$ is a suitable class of control with controlled state $X^\alpha = (X_t^\alpha)_{t \in \mathbb{N}}$ dynamics given by :

$$\begin{aligned} X^\alpha_{t+1} &= F(X_t, \alpha_t, \mathbb{P}^0_{(X_t, \alpha_t)}, \epsilon_{t+1}, \epsilon^0_{t+1}), \\ X^\alpha_0 &= \xi. \end{aligned} \tag{2}$$

where all the random variables are defined on an abstract filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

$\rightarrow$ The control problem (1)-(2) can be lifted on the space of measures $\mathcal{P}(\mathcal{X})$ and show that V is law invariant, ie for 2 $\mathcal{X}$-valued random variables ξ and ξ' satisfying $\mathbb{P}_\xi = \mathbb{P}_{\xi'}$, we have $V(\xi) = V(\xi')$.

$\rightarrow$ Extend the known CMKV-MDP theory to the case of non exchangeable interactions.
Non exchangeable interactions are motivated by recent litterature on Graphons.

- Graphon mean field systems :
  - Bayrakhtar, Chakraborty, Ruoyu Wu (22).
  - De Crescenzo, Coppini, Pham (23).
- Graphon mean field control (in continuous time):
  - Cao and Laurière (25).
  - De Crescenzo, Fuhrman, Kharroubi and Pham (24).
  - Kharroubi, Mekkaoui and Pham (25).

The agents labeled by $u \in I$ interact through a weighted probability measure through Graphons and functions of $\frac{\int_I G(u,v)\mathbb{P}_{X_t^v}(\mathrm{d}x)\mathrm{d}v}{\int_I G(u,v)\mathrm{d}v}$) where $G : I \times I \ni (u,v) \mapsto G(u,v)$ is a measurable map which measures the weight between agents u and v.

→ Extend the known CMKV-MDP theory to the case of non exchangeable interactions. Non exchangeable interactions are motivated by recent litterature on Graphons.

- Graphon mean field systems :
  - Bayrakhtar, Chakraborty, Ruoyu Wu (22).
  - De Crescenzo, Coppini, Pham (23).
- Graphon mean field control (in continuous time):
  - Cao and Laurière (25).
  - De Crescenzo, Fuhrman, Kharroubi and Pham (24).
  - Kharroubi, Mekkaoui and Pham (25).

The agents labeled by $u \in I$ interact through a weighted probability measure through Graphons and functions of $\frac{\int_I G(u,v)\mathbb{P}_{X_t^v}(\mathrm{d}x)\mathrm{d}v}{\int_I G(u,v)\mathrm{d}v}$) where $G : I \times I \ni (u, v) \mapsto G(u, v)$ is a measurable map which measures the weight between agents u and v.

→ We want to extend the framework of CMKV-MDP by introducing an adequate modelling of the heterogeneity between the agents.

N-agent formulation

- State dynamics for the controlled systems $\mathbf{X}^N = (X^{i,N})_{i\in 1,N}$

$$\begin{cases} X_0^{i,N} = x_0^i, \\ X_{t+1}^{i,N} = F_N\Big(\frac{i}{N}, X_t^{i,N}, \alpha_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} \delta_{(\frac{j}{N}, X_t^{j,N}, \alpha_t^{j,N})}, \epsilon_{t+1}^j, \epsilon_{t+1}^0\Big), \quad t \in \mathbb{N}. \end{cases} \tag{3}$$

- Value function for the N-agent system:

$$V_N^{\boldsymbol{\alpha}}(\boldsymbol{x}_0) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t f_N\Big(\frac{i}{N}, X_t^{i,N}, \alpha_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} \delta_{(\frac{j}{N}, X_t^{j,N}, \alpha_t^{j,N})}\Big)\Big], \tag{4}$$

where $\boldsymbol{x}_0 := (x_0^i)_{i\in 1,N} \in \mathcal{X}^N$ is the inital vector state of the agents. We then define

$$V_N(\boldsymbol{x}_0) := \sup_{\alpha\in\boldsymbol{\mathcal{A}}} V_N^{\alpha}(\boldsymbol{x}_0). \tag{5}$$

Strong and weak formulation for the non exchangeable mean field limit

- Strong formulation :

$$\begin{cases} X_0^u = \xi^u, \\ X_{t+1}^u = F(u, X_t^u, \alpha_t^u, \mathbb{P}^0_{(X_t^v, \alpha_t^v)}(\mathrm{d}x, \mathrm{d}a)\mathrm{d}v, \epsilon_{t+1}^u, \epsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad u \in I. \end{cases} \tag{6}$$

$$V^\alpha_{\text{strong}}(\boldsymbol{\xi}) := \int_I \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t f(u, X_t^u, \alpha_t^u, \mathbb{P}^0_{(X_t^v, \alpha_t^v)}(\mathrm{d}x, \mathrm{d}a)\mathrm{d}v)\Big]\mathrm{d}u, \quad V_{\text{strong}}(\boldsymbol{\xi}) := \sup_{\alpha\in\mathcal{A}} V^\alpha_{\text{strong}}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathcal{I}.$$

- Weak formulation :

$$\begin{cases} X_0 = \xi, \\ X_{t+1} = F(U, X_t, \alpha_t, \mathbb{P}^0_{(U, X_t, \alpha_t)}, \epsilon_{t+1}, \epsilon_{t+1}^0), \quad t \in \mathbb{N}. \end{cases} \tag{7}$$

$$V^\alpha_{\text{weak}}(\xi) := \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t f(U, X_t, \alpha_t, \mathbb{P}^0_{(U, X_t, \alpha_t)})\Big], \quad V_{\text{weak}}(\xi) := \sup_{\alpha\in\mathcal{A}} V^\alpha_{\text{weak}}(\xi), \quad \xi \in \mathcal{I}. \tag{8}$$

where U is a uniform random variable $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ encoding the heterogeneity.

We will work under the weak formulation and show further a connection with the strong formulation.

Objectives :

- Show how the control problem (7)- (8) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \left\{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\right\} \tag{9}$$

where $\mathrm{pr}_1 : I \times \mathcal{X} \ni (u, x) \mapsto \mathrm{pr}_1(u, x) = u$ and $\#$ is the pushforward notation. We will then characterize the value function V_{weak} as a fixed point of a suitable Bellman operator on $\mathcal{P}_\lambda(I \times \mathcal{X})$.

We will work under the weak formulation and show further a connection with the strong formulation.

Objectives :

- Show how the control problem (7)- (8) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\} \tag{9}$$

  where $\mathrm{pr}_1 : I \times \mathcal{X} \ni (u, x) \mapsto \mathrm{pr}_1(u, x) = u$ and $\#$ is the pushforward notation. We will then characterize the value function V_{weak} as a fixed point of a suitable Bellman operator on $\mathcal{P}_\lambda(I \times \mathcal{X})$.
- Show a quantitative propagation of chaos for the convergence of the value function of the N-agent MDP V_N defined in (5) towards V_{weak} and V_{strong} for all $\boldsymbol{x} := (x^i)_{i \in \{1,N\}}$ satisfying a regularity condition to be precised later and show how to construct approximate optimal policies for the N-agent MDP from optimal randomized feedback control of the CNEMF-MDP.

We will work under the weak formulation and show further a connection with the strong formulation.

Objectives :

- Show how the control problem (7)- (8) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \left\{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\right\} \tag{9}$$

  where $\mathrm{pr}_1 : I \times \mathcal{X} \ni (u, x) \mapsto \mathrm{pr}_1(u, x) = u$ and $\#$ is the pushforward notation. We will then characterize the value function V_{weak} as a fixed point of a suitable Bellman operator on $\mathcal{P}_\lambda(I \times \mathcal{X})$.
- Show a quantitative propagation of chaos for the convergence of the value function of the N-agent MDP V_N defined in (5) towards V_{weak} and V_{strong} for all $\boldsymbol{x} := (x^i)_{i \in \{1,N\}}$ satisfying a regularity condition to be precised later and show how to construct approximate optimal policies for the N-agent MDP from optimal randomized feedback control of the CNEMF-MDP.
- Propose a simple application of our non exchangeable mean field model to the case of targeting advertising.

Contents

1. Introduction
   - Context and motivations
   - Problem formulation

2. Bellman fixed point on the space $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Lifted MDP on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Bellman fixed point on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Equivalence with the strong formulation

3. Propagation of chaos from V_N towards V
   - The N-agent problem as a MDP on the state space $\mathcal{X}^N$ and action space A^N
   - Convergence of value functions
   - Approximate optimal policies

Regularity assumptions on f and F

- Regularity on the state transition function F

$$\mathbb{E}\Big[d\big(F(u,x,a,\mu,\epsilon_1^1,e^0),F(u,x',a,\mu',\epsilon_1^1,e^0)\big)\Big] \leqslant L_F\big(\mathrm{d}(x,x')+\mathcal{W}(\mu,\mu')\big). \quad (10)$$

- Regularity on the reward function f

$$\big|f(u,x,a,\mu)-f(u,x',a,\mu')\big| \leqslant L_f\big(\mathrm{d}(x,x')+\mathcal{W}(\mu,\mu')\big). \quad (11)$$

for every $u \in I$, $x,x' \in \mathcal{X}$, $a \in A$, $\mu,\mu' \in \mathcal{P}(I \times \mathcal{X} \times A)$ and $e^0 \in E^0$.

- The Lipschitz assumption on F is made on expectation, and not pathwisely.
- The definition of the mean-field limit doesn't require any regularity assumption on the label u.

Lifting the MDP on $\mathcal{P}_\lambda(I \times \mathcal{X})$

Define the measurable map $\tilde{F} : I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \times E \times E^0 \to I \times \mathcal{X}$ as

$$\tilde{F}(u, x, a, \mu, e, e^0) = \big(u, F(u, x, a, \mu, e, e^0)\big).$$

- Set $\mu_{t+1} = \mathbb{P}^0_{(U, X_{t+1})} \in \mathcal{P}_\lambda(I \times \mathcal{X})$. Then (using the pushforward notation $\#$):

$$\mu_{t+1} = \tilde{F}(., ., .\mathbb{P}^0_{(U, X_t, \alpha_t)}, \epsilon^0_{t+1}) \# (\mathbb{P}^0_{(U, X_t, \alpha_t)} \otimes \lambda_\epsilon) \quad \mathbb{P}\text{-a.s}, \tag{12}$$

Bayes Formula gives $\mathbb{P}^0_{(U, X_t, \alpha_t)} = \mu_t \hat{\otimes} \hat{\alpha}_t$ where $\hat{\alpha}_t$ is a probability kernel:

$$\hat{\alpha}_t : I \times \mathcal{X} \ni (u, x) \mapsto \mathbb{P}^0_{\alpha_t | (U, X_t) = (u, x)} \in \mathcal{P}(A), \tag{13}$$

$$\mu_{t+1} = \hat{F}(\mu_t, \hat{\alpha}_t, \epsilon^0_{t+1}), \quad t \in \mathbb{N}, \tag{14}$$

with $\hat{F}(\hat{\mu}, \hat{a}, e^0) := \tilde{F}(., ., ., \hat{\mu} \hat{\otimes} \hat{a}, ., e^0) \# ((\hat{\mu} \hat{\otimes} \hat{a}) \otimes \lambda_\epsilon)$. and relaxed ($\mathcal{P}(A)$-valued) feedback control $\hat{\alpha}$ on $I \times \mathcal{X}$.

- Similarly and with law of conditional expectations, we have

$$V^\alpha = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \hat{f}(\mu_t, \hat{\alpha}_t)\Big], \tag{15}$$

for some measurable function $\hat{f} : \mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0(I \times \mathcal{X}; \mathcal{P}(A)) \to \mathbb{R}$ explicitly derived from f.

Definition of the Bellman operator $\mathcal{T}$

- operator $\mathcal{T}$ of the lifted MDP: For $\mathcal{W} \in L^\infty_m(\mathcal{P}_\lambda(I \times \mathcal{X}))$,

$$\big[\mathcal{T} W\big](\mu) = \sup_{\hat{a} \in L^0(I \times \mathcal{X}; \mathcal{P}(A))} \big[\hat{\mathcal{T}}^{\hat{a}} W\big](\mu) = \sup_{a \in L^0(I \times \mathcal{X} \times [0,1]; A)} \big[\mathbb{T}^a W\big](\mu), \tag{16}$$

where $\hat{\mathcal{T}}^{\hat{a}}$ and $\mathbb{T}^a$ are operators defined on $L^\infty\big(\mathcal{P}_\lambda(I \times \mathcal{X})\big)$ by

$$\begin{aligned}
\big[\hat{\mathcal{T}}^{\hat{a}} W\big](\mu) &:= \hat{f}(\mu, \hat{a}) + \beta \mathbb{E}\big[W(\hat{F}(\mu, \hat{a}, \epsilon_1^0))\big] \\
\big[\mathbb{T}^a W\big](\mu) &:= \mathbb{E}\Big[f(\xi, a(\xi, U), \mathcal{L}\big(\xi, a(\xi, U)\big) + \beta W\Big(\mathbb{P}^0_{\tilde{F}(\xi, a(\xi, U), \mathcal{L}(\xi, a(\xi, U), \epsilon_1, \epsilon_1^0)}\Big)\Big],
\end{aligned} \tag{17}$$

for any $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$.

Theorem

- **Law invariance.** For any ξ and ξ' $\mathcal{X}$-valued random variables s.t $\mathbb{P}_{(U,\xi)} = \mathbb{P}_{(U,\xi')}$, we have $V_{\text{weak}}(\xi) = V_{\text{weak}}(\xi')$. We then define $V(\mu) := V_{\text{weak}}(\xi)$, for $\mu = \mathbb{P}_{(U,\xi)} \in \mathcal{P}_\lambda(I \times \mathcal{X})$.

Characterization by Bellman equation on $\mathcal{P}_\lambda(I \times \mathcal{X})$

Theorem

- **Law invariance.** For any ξ and ξ' $\mathcal{X}$-valued random variables s.t $\mathbb{P}_{(U,\xi)} = \mathbb{P}_{(U,\xi')}$, we have $V_{\text{weak}}(\xi) = V_{\text{weak}}(\xi')$. We then define $V(\mu) := V_{\text{weak}}(\xi)$, for $\mu = \mathbb{P}_{(U,\xi)} \in \mathcal{P}_\lambda(I \times \mathcal{X})$.
- **Dynamic Programming.** We have V_{weak} fixed point for the operator $\mathcal{T}$:

$$V_{\text{weak}}(\mu) = \big[\mathcal{T} V_{\text{weak}}\big](\mu), \quad \mu \in \mathcal{P}_\lambda(I \times \mathcal{X}) \tag{18}$$

- **Existence of optimal randomized feedback control α^* for $V_{\text{weak}}(\xi)$** in the form:

Characterization by Bellman equation on $\mathcal{P}_\lambda(I \times \mathcal{X})$

Theorem

- **Law invariance.** For any ξ and ξ' $\mathcal{X}$-valued random variables s.t $\mathbb{P}_{(U,\xi)} = \mathbb{P}_{(U,\xi')}$, we have $V_{\text{weak}}(\xi) = V_{\text{weak}}(\xi')$. We then define $V(\mu) := V_{\text{weak}}(\xi)$, for $\mu = \mathbb{P}_{(U,\xi)} \in \mathcal{P}_\lambda(I \times \mathcal{X})$.
- **Dynamic Programming.** We have V_{weak} fixed point for the operator $\mathcal{T}$:

$$V_{\text{weak}}(\mu) = \big[\mathcal{T} V_{\text{weak}}\big](\mu), \quad \mu \in \mathcal{P}_\lambda(I \times \mathcal{X}) \tag{18}$$

- **Existence of optimal randomized feedback control α^* for $V_{\text{weak}}(\xi)$** in the form:

$$\alpha_t^* = a^*(\mathbb{P}^0_{(U,X_t)}, U, X_t, V_t) \tag{19}$$

  where $(V_t)_{t\in\mathbb{N}}$ sequence of $i.i.d$ uniform random variables for some measurable function $a^*(\mu, u, x, \tilde{u})$ on $\mathcal{P}_\lambda(I \times \mathcal{X}) \times I \times \mathcal{X} \times [0, 1]$.

Formulation of the strong formulation

- State dynamics for the controlled systems $\mathbf{X} = (X^u)_{u \in I}$:

$$\begin{cases} X_0^u = \xi^u, \\ X_{t+1}^u = F(u, X_t^u, \alpha_t^u, \mathbb{P}^0_{(X_t^v, \alpha_t^v)}(\mathrm{d}x, \mathrm{d}a)\mathrm{d}v, \epsilon_{t+1}^u, \epsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad u \in I. \end{cases} \tag{20}$$

- Value function in the strong formulation :

$$V_{\text{strong}}^{\alpha}(\boldsymbol{\xi}) := \int_I \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t f(u, X_t^u, \alpha_t^u, \mathbb{P}^0_{(X_t^v, \alpha_t^v)}(\mathrm{d}x, \mathrm{d}a)\mathrm{d}v)\Big]\mathrm{d}u, \quad \boldsymbol{\xi} = (\xi^u)_{u \in I}. \tag{21}$$

  The value function of the conditional non exchangeable mean field control Markov decision processes CNEMF-MDP is then defined by

$$V_{\text{strong}}(\boldsymbol{\xi}) := \sup_{\alpha \in \mathcal{A}} V_{\text{strong}}^{\alpha}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathcal{I}. \tag{22}$$

- Note that the uncountable collection of *i.i.d* random variables $(\epsilon^u)_{u \in I}$ induces some measurability issues for the formulation of the strong formulation compared to the weak formulation.

Proposition (Equivalence of value functions).

Let $\boldsymbol{\xi} = (\xi^u)_{u\in I}$ and ξ be random variables such that $\mathbb{P}_{\xi^u} = \mathbb{P}_{\xi|U=u}$ for λ a.e $u \in I$. Then , we have

$$V_{\text{strong}}(\boldsymbol{\xi}) = V_{\text{weak}}(\xi) = V(\mu), \quad \mu = \mathbb{P}_{(U,\xi)} = \mathbb{P}_{\xi^u}(\mathrm{d}x)\mathrm{d}u. \tag{23}$$

Proof.

$\square$

$\to$ We now denote indifferently V to denote V_{strong} or V_{weak}.

Contents

1. Introduction
   - Context and motivations
   - Problem formulation
2. Bellman fixed point on the space $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Lifted MDP on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Bellman fixed point on $\mathcal{P}_\lambda(I \times \mathcal{X})$
   - Equivalence with the strong formulation
3. Propagation of chaos from V_N towards V
   - The N-agent problem as a MDP on the state space $\mathcal{X}^N$ and action space A^N
   - Convergence of value functions
   - Approximate optimal policies

The N-agent problem as a MDP on state space $\mathcal{X}^N$ and action space A^N.

Formulation of the N-agent MDP

- State dynamics for the N-agent controlled systems $\mathbf{X}^N = (X_i^N)_{i \in 1,N}$

$$\begin{cases} X_0^i = x_0^i, \\ X_{t+1}^i = F_N\left(\frac{i}{N}, X_t^i, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(\frac{j}{N}, X_t^j, \alpha_t^j)}, \epsilon_{t+1}^j, \epsilon_{t+1}^0\right), \quad t \in \mathbb{N}. \end{cases} \tag{24}$$

where $\boldsymbol{x}_0 := (x_0^i)_{i \in 1,N} \in \mathcal{X}^N$ is the inital vector state of the agents.

- Value function for the N agent MDP.

$$V_N^{\boldsymbol{\alpha}}(\boldsymbol{x}_0) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t f_N\Big(\frac{i}{N}, X_t^i, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(\frac{j}{N}, X_t^j, \alpha_t^j)}\Big)\Big], \tag{25}$$

$$V_N(\boldsymbol{x}_0) := \sup_{\alpha \in \boldsymbol{\mathcal{A}}} V_N^{\alpha}(\boldsymbol{x}_0). \tag{26}$$

The N-agent problem as a MDP on the space $\mathcal{X}^N$.

MDP on the space $\mathcal{X}^N$.

- State dynamics (24) can be written :

$$\mathbf{X}_{t+1} = \boldsymbol{F}_N(\mathbf{X}_t, \boldsymbol{\alpha}_t, \boldsymbol{\epsilon}_{t+1}), \tag{27}$$

with state transition function $\boldsymbol{F}_N : \mathcal{X}^N \times A^N \times (E^N \times E^0) \to \mathcal{X}^N$ is given for $\boldsymbol{x} = (x^i)_{i \in 1,N}$, $= (a^i)_{i \in 1,N}$ and $\mathbf{e} = \big((e^i)_{i \in 1,N}, e^0\big)$ by

$$\boldsymbol{F}_N(\boldsymbol{x}, \boldsymbol{a}, \mathbf{e}) := \Big(F_N(\frac{i}{N}, x^i, a^i, \frac{1}{N}\sum_{i=1}^N \delta_{(\frac{i}{N}, x^i, a^i)}, e^i, e^0)\Big)_{i \in 1,N},$$

- Value function (25) for the N agent MDP :

$$V_N^{\alpha}(\boldsymbol{x}_0) = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \boldsymbol{f}_N(\mathbf{X}_t, \boldsymbol{\alpha}_t)\Big]. \tag{28}$$

with reward function $\boldsymbol{f}_N : \mathcal{X}^N \times A^N \to \mathbb{R}$ is given by

$$\boldsymbol{f}_N(\boldsymbol{x}, \boldsymbol{a}) := \frac{1}{N}\sum_{i=1}^N f_N(\frac{i}{N}, x^i, a^i, \frac{1}{N}\sum_{i=1}^N \delta_{(\frac{i}{N}, x^i, a^i)}), \quad \boldsymbol{x} = (x^i)_{i \in 1,N}, \quad \boldsymbol{a} = (a^i)_{i \in 1,N}.$$

the N-agent problem as a MDP on $\mathcal{X}^N$ with control space A^N

N-agent MDP formulation

- Bellman operator for the N-agent MDP

$$\big[\mathcal{T}_N W\big](\boldsymbol{x}) := \sup_{\boldsymbol{a} \in A^N} \mathbb{T}_N^{\boldsymbol{a}} W(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}^N. \tag{29}$$

where

$$\mathbb{T}_N^{\boldsymbol{a}} W(\boldsymbol{x}) := f_N(\boldsymbol{x}, \boldsymbol{a}) + \beta \mathbb{E}\Big[W\big(\boldsymbol{F}_N(\boldsymbol{x}, \boldsymbol{a}, \epsilon_1)\big)\Big], \quad \boldsymbol{x} \in \mathcal{X}^N, \quad \boldsymbol{a} \in A^N. \tag{30}$$

Assumption on the regularity of the initial condition

For a given $\boldsymbol{x} := (x^1, x^2, \ldots, x^N) \in \mathcal{X}^N$, we say that $\boldsymbol{x}$ is regular if the following condition holds true. There exists a constant $C > 0$ such that for any $i, j \in \{1, \ldots, N\}$,

$$d(x^i, x^j) \leqslant C \frac{|i-j|}{N}. \tag{31}$$

The set of regular $\boldsymbol{x}$ will be denoted in the following $\mathcal{X}^N_{\text{reg}}$.

- The assumption (31) is crucial in the derivation of the propagation of chaos result

Assumption on the regularity of f and F with respect to the label state

(i) Let $N \in \mathbb{N}^*$. The mapping

$$I \ni u \mapsto f(u, x, a, \mu) \in \mathbb{R}, \tag{32}$$

has a bounded variation on the interval $[\frac{j-1}{N}, \frac{j}{N}[$ which we denoted by $V_{\frac{j-1}{N}}^{\frac{j}{N}}(f)$ (by omitting the dependance in (x, a, μ) which satisfies

$$V_{\frac{j-1}{N}}^{\frac{j}{N}}(f) \leqslant \frac{C}{N} \text{ or } \frac{C}{\sqrt{N}}, \tag{33}$$

for every $j \in \{1, \ldots, N\}$ and for every $(x, a, \mu) \in \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A)$.

(ii)

$$\mathbb{E}\big[d\big(F(u, x, a, \mu, \epsilon_1^1, e^0), F(u', x', a', \mu', \epsilon_1^1, e^0)\big)\big] \leqslant K_F\big(\mathbf{d}((u, x, a), (u', x', a')) + \mathcal{W}(\mu, \mu')\big), \tag{34}$$

for every $u, u' \in I$, $x, x' \in \mathcal{X}$, $a, a' \in A$ and $\mu, \mu' \in \mathcal{P}(I \times \mathcal{X} \times A)$, $e^0 \in E^0$.

Propagation of chaos of value functions

Theorem : Convergence of value functions and propagation of chaos

- For V value function on $\mathcal{P}_\lambda(I \times \mathcal{X})$ of the CNEMF-MDP, we set the lifted operator $\widetilde{V}$ defined on $\mathcal{X}^N$ by

$$\widetilde{V}(\boldsymbol{x}) := V(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}]), \quad \text{for } \boldsymbol{x} = (x^i)_{i \in \{1,N\}} \in \mathcal{X}^N, \tag{35}$$

  where $\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}] := \sum_{j=1}^N \mathbb{1}_{[\frac{j-1}{N}, \frac{j}{N}[}(u)\delta_{x^j}(\mathrm{d}x)\mathrm{d}u \in \mathcal{P}_\lambda(I \times \mathcal{X})$.

- There exists some positive constant C such that for all $\boldsymbol{x} := (x^i)_{i \in \{1,N\}} \in \mathcal{X}^N_{\text{reg}}$, we have

$$\left|V_N(\boldsymbol{x}) - V(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}])\right| \underset{N \to \infty}{\longrightarrow} 0. \tag{36}$$

  Moreover, propagation of chaos rate of convergence takes the following form

$$\left|V_N(\boldsymbol{x}) - V(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}])\right| \leqslant C\Big({M_N}^\gamma + O(N^{-\frac{\gamma}{2}}) + \|f - f^N\|_\infty + \|F - F^N\|_\infty^\gamma + \frac{1}{N}\sum_{j=1}^N V_{\frac{j-1}{N}}^{\frac{j}{N}}(f)\Big). \tag{37}$$

  with $M_N := \sup_{\nu \in \mathcal{P}(I \times \mathcal{X} \times A)} \mathbb{E}\Big[\mathcal{W}(\nu_N, \nu)\Big]$, $\quad$ (ν_N empirical measure of ν).

It extends the result from [1] with the additional errors:

- $O(N^{-\frac{\gamma}{2}})$ which represents the error due to the label convergence.
- $\|f - f_N\|_\infty$ and $\|F - F_N\|_\infty$ which represent the errors due to the convergence of the state dynamics functions and the reward functions.

Theorem : Approximate optimal policies

Let $\mathfrak{a}^* : \mathcal{P}_\lambda(I \times \mathcal{X}) \times I \times \mathcal{X} \times [0,1] \to A$ be an optimal randomized feedback policy for the CNEMF-MDP. Then, defining

$$\pi_r^{\mathfrak{a}^*,N}(\boldsymbol{x}, \boldsymbol{u}) := \left(\mathfrak{a}^*(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}], \frac{i}{N}, x^i, u^i\right)_{i \in \{1,N\}}, \tag{38}$$

for $\boldsymbol{x} := (x^i)_{i\in\{1,N\}} \in \mathcal{X}_{\text{reg}}^N$, $\boldsymbol{u} = (u^i)_{i\in\{1,N\}}$. Then, the randomized feedback control $\boldsymbol{\alpha}_t^{r,,N} \in \mathcal{A}$ defined as

$$\boldsymbol{\alpha}_t^{r,N} = \pi_r^{\mathfrak{a}^*,N}(\mathbf{X}_t, \boldsymbol{U}_t), \quad t \in \mathbb{N}, \tag{39}$$

where $\left\{\boldsymbol{U}_t = (U_t^i)_{i\in\{1,N\}}, t \in \mathbb{N}\right\}$ is a family of mutually i.i.d uniform random variables on $[0,1]$, is an $O(M_N^\gamma + N^{-\frac{\gamma}{2}} + \|f - f^N\|_\infty + \|F - F^N\|_\infty^\gamma)$ optimal control for the N-agent MDP.

Conclusion of our work

- CNEMF-MDP lifted to optimization problem on the space $\mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X})$ with relaxed controls valued in $\boldsymbol{A} = \mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X} \times A)$ with marginal constraint $\rightarrow$ Standard MFC on the Wasserstein space $\mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X})$.
  - Characterization of the value function as a fixed point of a Bellman operator.
  - Equivalence formulation between weak and strong formulation.
  - Existence of an optimal randomized feedback control policy a^*.

Conclusion of our work

- CNEMF-MDP lifted to optimization problem on the space $\mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X})$ with relaxed controls valued in $\boldsymbol{A} = \mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X} \times A)$ with marginal constraint → Standard MFC on the Wasserstein space $\mathcal{P}_\lambda(\boldsymbol{I} \times \mathcal{X})$.
  - Characterization of the value function as a fixed point of a Bellman operator.
  - Equivalence formulation between weak and strong formulation.
  - Existence of an optimal randomized feedback control policy a^*.
- Convergence of the value function V_N of the N-agent MDP towards the value function V for initial state agents $\boldsymbol{x} := (x^i)_{i \in \{1,\ldots,N\}} \in \mathcal{X}^N_{\text{reg}}$ with explicit rate of convergence

Conclusion of our work

- CNEMF-MDP lifted to optimization problem on the space $\mathcal{P}_\lambda(I \times \mathcal{X})$ with relaxed controls valued in $\boldsymbol{A} = \mathcal{P}_\lambda(I \times \mathcal{X} \times A)$ with marginal constraint $\rightarrow$ Standard MFC on the Wasserstein space $\mathcal{P}_\lambda(I \times \mathcal{X})$.
  - Characterization of the value function as a fixed point of a Bellman operator.
  - Equivalence formulation between weak and strong formulation.
  - Existence of an optimal randomized feedback control policy a^*.
- Convergence of the value function V_N of the N-agent MDP towards the value function V for initial state agents $\boldsymbol{x} := (x^i)_{i \in \{1,\dots,N\}} \in \mathcal{X}^N_{\text{reg}}$ with explicit rate of convergence
- Optimal randomized feedback control for CNEMF-MDP $\rightarrow$ Quantitative approximate optimal policy for the N-agent MDP.

Future works on non exchangeable mean field systems

- Numerical algorithms in the context of a finite number of players :
  (1) In a model-based setting : Learning optimal controls $\boldsymbol{\alpha} = (\alpha^{1,N}, \ldots, \alpha^{N,N})$ and value function V_N through Deep Learning algorithms.
  (2) In a model-free setting : Learning optimal policies and value function V_N through Reinforcement Learning algorithms.

Future works on non exchangeable mean field systems

- Numerical algorithms in the context of a finite number of players :
  (1) In a model-based setting : Learning optimal controls $\boldsymbol{\alpha} = (\alpha^{1,N}, \ldots, \alpha^{N,N})$ and value function V_N through Deep Learning algorithms.
  (2) In a model-free setting : Learning optimal policies and value function V_N through Reinforcement Learning algorithms.
- Numerical algorithms in the context of the label state formulation
  (1) In a model-based setting : Learning optimal control $\boldsymbol{\alpha}$ through Deep Learning algorithm and value function V through HJB equation
  (2) In a model-free setting : Learning optimal policy π and optimal value function V through Actor-critic algorithms.

Future works on non exchangeable mean field systems

- Numerical algorithms in the context of a finite number of players :
  (1) In a model-based setting : Learning optimal controls $\boldsymbol{\alpha} = (\alpha^{1,N}, \ldots, \alpha^{N,N})$ and value function V_N through Deep Learning algorithms.
  (2) In a model-free setting : Learning optimal policies and value function V_N through Reinforcement Learning algorithms.
- Numerical algorithms in the context of the label state formulation
  (1) In a model-based setting : Learning optimal control $\boldsymbol{\alpha}$ through Deep Learning algorithm and value function V through HJB equation
  (2) In a model-free setting : Learning optimal policy π and optimal value function V through Actor-critic algorithms.
- LQ control problem for non exchangeable mean field systems (with common noise).

References

M. Motte and H. Pham. *Mean-field Markov decision processes with common noise and open-loop controls, The Annals of Applied Probability, 32(2):1421–1458, 2022.*

M. Motte and H. Pham. *Quantitative propagation of chaos for mean field Markov decision process with common noise.* Electronic Journal of Probability, 28:1–24, 2023.

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

S.Mekkaoui, H.Pham *Analysis of Non-Exchangeable Mean Field Markov Decision Processes with common noise : From Bellman equation to quantitative propagation of chaos. Work in Progress*

Thank you for your attention