# 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

## 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: $(\epsilon_t^u)_{u \in I, t \in \mathbb{N}}$, 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 $I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \to \mathbb{R}$.

## 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 $X^{\alpha} = (X^{\alpha}_t)_{t \in \mathbb{N}}$ 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 $\xi$ and $\xi'$ 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 in the form $\frac{\int_I G(u,v)\mathbb{P}_{X_t^v}(\mathrm{d}x)\mathrm{d}v}{\int_I G(u,v)\mathrm{d}v})$ where $G : I \times I \ni (u, v) \mapsto G(u, v)$ is a measurable map which measures the weight between agents $u$ and $v$.

$\rightarrow$ 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 in the form $\frac{\int_I G(u,v)\mathbb{P}_{X_t^v}(\mathrm{d}x)\mathrm{d}v}{\int_I G(u,v)\mathrm{d}v})$ where $G : I \times I \ni (u,v) \mapsto G(u,v)$ is a measurable map which measures the weight between agents $u$ and $v$.

$\rightarrow$ 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 $\mathbf{X}^N = (X^{i,N})_{i\in[\![1,N]\!]}$

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

- Value function for the $N$-agent system:

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

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

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

where

$$\boldsymbol{\mathcal{A}} := \Big\{\boldsymbol{\alpha} = (\alpha_t^i)_{i\in 1,N, t\in\mathbb{N}} : \alpha^i \text{ is } \mathbb{F}^N\text{-adapted for each } i\in[\![1,N]\!] \Big\},$$

and where $\mathbb{F}^N := (\mathcal{F}_t^N)_{t\in\mathbb{N}}$ generated by $\boldsymbol{\epsilon}^N = ((\epsilon_t^i)_{i\in[\![1,N]\!]}, \epsilon_t^0)_{t\in\mathbb{N}^\star}$ completed with a family of mutually i.i.d uniform random variables $\boldsymbol{U}^N = (U_t^i)_{i\in[\![1,N]\!],\, t\in\mathbb{N}}$ used for randomizing the controls $(\alpha^i)_{i\in[\![1,N]\!]}$.

## Strong formulation for the non exchangeable mean field limit

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

$$\begin{cases} X_0^u = \xi^u, \\ X_{t+1}^u = F(u, X_t^u, \alpha_t^u, \mathbb{P}^0_{(X_t^v, \alpha_t^v)}(\mathrm{d}x, \mathrm{d}a)\mathrm{d}v, \epsilon_{t+1}^u, \epsilon_{t+1}^0), \quad t \in \mathbb{N}, \quad u \in I. \end{cases} \tag{3}$$

- Value function in the strong formulation:

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

where $\boldsymbol{\xi} = (\xi^u)_{u \in I}$ denotes the collection of random initial values and

$$\mathcal{A}^{\text{strong}} := \big\{\boldsymbol{\alpha} = (\alpha_t^u)_{u \in I, t \in \mathbb{N}} : \alpha_t^u = \alpha_t(u, \Gamma^u, (\epsilon_s^u)_{s \leqslant t}, (\epsilon_s^0)_{s \leqslant t}) \text{ for every } t \in \mathbb{N}\big\}.$$

where
→ $\Gamma^u$ denotes the initial information available for agent $u$ supposed to admit an extra random variable $U^u \sim \mathcal{U}([0,1])$ independant of $\xi^u$ and $\mathcal{G}^u = \sigma(\Gamma^u)$-measurable.
→ $\mathcal{I}$ denotes an admissible class of initial conditions ensuring the measurability of $u \mapsto \mathbb{P}_{(X_t^u, \alpha_t^u, (\epsilon_s^0)_{s \leqslant t})}$ for any $t \in \mathbb{N}$ , hence the well posedness of the cost functional

## Weak formulation for the non exchangeable mean field limit

- State dynamics for the controlled system $X$

$$\begin{cases} X_0 = \xi, \\ X_{t+1} = F(U, X_t, \alpha_t, \mathbb{P}^0_{(U,X_t,\alpha_t)}, \epsilon_{t+1}, \epsilon^0_{t+1}), \quad t \in \mathbb{N}. \end{cases} \tag{4}$$

- Value function in the weak formulation:

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

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

$$\mathcal{A}^{\text{weak}} := \big\{\boldsymbol{\alpha} = (\alpha_t)_{t\in\mathbb{N}} : \alpha_t = \alpha_t(U, \Gamma, (\epsilon_s)_{s\leqslant t}, (\epsilon^0_s)_{s\leqslant t}) \text{ for every } t \in \mathbb{N}\big\}.$$

$\rightarrow$ The weak formulation (4) should be understood as a relaxed formulation of (3) which avoids measurability issues but lacks of pathwise interpretation.

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

## Objectives :

- Show how the control problem (4)- (5) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \left\{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\right\} \quad (5)$$

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 $V_{\mathrm{weak}}$ 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 (4)- (5) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\} \tag{5}$$

  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 $V_{\mathrm{weak}}$ 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 $V_N$ towards $V_{\mathrm{weak}}$ and $V_{\mathrm{strong}}$ 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 (4)- (5) called CNEMF-MDP can be recasted as a standard mean field control problem on the space

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \left\{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\right\} \tag{5}$$

  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 $V_{\text{weak}}$ 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 $V_N$ towards $V_{\text{weak}}$ and $V_{\text{strong}}$ 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

## Regularity assumptions on $f$ and $F$

- Regularity on the state transition function $F$

$$\mathbb{E}\Big[d\Big(F(u,x,a,\mu,\epsilon_1^1,e^0),F(u,x',a,\mu',\epsilon_1^1,e^0)\Big)\Big] \leqslant L_F\big(\mathrm{d}(x,x') + \mathcal{W}(\mu,\mu')\big). \tag{6}$$

- Regularity on the reward function $f$

$$\big|f(u,x,a,\mu) - f(u,x',a,\mu')\big| \leqslant L_f\big(\mathrm{d}(x,x') + \mathcal{W}(\mu,\mu')\big). \tag{7}$$

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 $\#$):

$$\mu_{t+1} = \tilde{F}(\cdot, \cdot, \cdot, \mathbb{P}^0_{(U, X_t, \alpha_t)}, \cdot, \epsilon^0_{t+1}) \# \big(\mathbb{P}^0_{(U, X_t, \alpha_t)} \otimes \lambda_\epsilon\big) \quad \mathbb{P}\text{-a.s}, \quad t \in \mathbb{N}. \tag{8}$$

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

## 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 $\#$):

$$\mu_{t+1} = \tilde{F}(\cdot, \cdot, \cdot, \mathbb{P}^0_{(U, X_t, \alpha_t)}, \cdot, \epsilon^0_{t+1}) \# \big(\mathbb{P}^0_{(U, X_t, \alpha_t)} \otimes \lambda_\epsilon\big) \quad \mathbb{P}\text{-a.s}, \quad t \in \mathbb{N}. \tag{8}$$

Considering the $\mathbb{F}^0$-adapted control process $\boldsymbol{\alpha}_t = \mathbb{P}^0_{(U, X_t, \alpha_t)}$ (Note that this process has to satisfy $\mathrm{pr}_{12} \# \boldsymbol{\alpha}_t = \mu_t$), and from a suitable measurable coupling ensuring that one can find a measurable map

$$\mathbf{p} : \mathcal{P}_\lambda(I \times \mathcal{X}) \times \mathcal{P}_\lambda(I \times \mathcal{X} \times A) \to \mathcal{P}_\lambda(I \times \mathcal{X} \times A),$$

such that $\mathrm{pr}_{12} \# \mathbf{p}(\mu, \boldsymbol{a}) = \mu$ and if $\mathrm{pr}_{12} \# \boldsymbol{a} = \mu$, then $\mathbf{p}(\mu, \boldsymbol{a}) =$. It follows that (11) can be rewritten as

$$\mu_{t+1} = \hat{F}\big(\mu_t, \boldsymbol{\alpha}_t, \epsilon^0_{t+1}\big), \qquad \mathbb{P}\text{-a.s}, \quad t \in \mathbb{N}, \tag{9}$$

with $\hat{F}(\mu, \boldsymbol{a}, e^0) := \tilde{F}\big(\cdot, \cdot, \cdot, \mathbf{p}(\mu, \boldsymbol{a}) . \cdot, e^0\big) \# \big((\mathbf{p}(\mu, \boldsymbol{a}) \otimes \lambda_\epsilon\big).$

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

- Similarly and with law of conditional expectations,

$$V^{\alpha}_{\text{weak}}(\xi) = \hat{V}^{\alpha}(\mu_0) = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \hat{f}(\mu_t, \boldsymbol{\alpha}_t)\Big], \quad \text{with } \mu_0 = \mathbb{P}_{(U,\xi)} \in \mathcal{P}_\lambda(I \times \mathcal{X}). \tag{10}$$

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

$$\hat{f}(\mu, \boldsymbol{a}) := \int_{I \times \mathcal{X} \times A} f\big(u, x, a, \boldsymbol{p}(\mu, \boldsymbol{a})\big)\boldsymbol{p}(\mu, \boldsymbol{a})(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a).$$

Defining $\boldsymbol{\mathcal{A}}$ as the set of $\mathbb{F}^0-$adapted processed valued in $\boldsymbol{A} = \mathcal{P}_\lambda(I \times \mathcal{X} \times A)$ and denoting $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we define

$$\begin{cases} \hat{V}^{\boldsymbol{\nu}}(\mu_0) & = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \hat{f}(\mu_t, \boldsymbol{\nu}_t)\Big], \\ \hat{V}(\mu_0) & = \sup\limits_{\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}} \hat{V}^{\boldsymbol{\alpha}}(\mu_0). \end{cases} \tag{11}$$

with dynamics $\mu_{t+1} = \hat{F}(\mu_t, \nu_t, \epsilon^0_{t+1})$.

$\rightarrow$ From (10), we can see that $V_{\text{weak}}(\xi) \leqslant \hat{V}(\mu)$ when $\mu = \mathbb{P}_{(U,\xi)}$, and the goal is to show the equality.

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

- Bellman operator $\mathcal{T}$ defined on $L^\infty_m(\mathcal{P}_\lambda(I \times \mathcal{X}))$ = bounded $\mathbb{R}$-valued measurable maps on $\mathcal{P}_\lambda(I \times \mathcal{X}))$.

$$[\mathcal{T} W](\mu) := \sup_{\boldsymbol{a} \in \boldsymbol{A}} \big\{ \hat{f}(\mu, \boldsymbol{a}) + \beta \mathbb{E}\big[ W(\hat{F}(\mu, \boldsymbol{a}, \epsilon_1^0)) \big] \big\}, \quad \mu \in \mathcal{P}_\lambda(I \times \mathcal{X}).$$

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

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

where $\mathbb{T}^a$ is an operator defined on $L^\infty\big(\mathcal{P}_\lambda(I \times \mathcal{X})\big)$ by

$$\big[\mathbb{T}^a W\big](\mu) := \mathbb{E}\Big[ f(\boldsymbol{\xi}, \mathrm{a}(\boldsymbol{\xi}, \tilde{U}), \mathbb{P}_{\left(\boldsymbol{\xi}, \mathrm{a}(\boldsymbol{\xi}, \tilde{U})\right)} + \beta W\big( \mathbb{P}^0_{\tilde{F}(\boldsymbol{\xi}, a(\boldsymbol{\xi}, \tilde{U}), \mathbb{P}_{(\boldsymbol{\xi}, \mathrm{a}(\boldsymbol{\xi}, \tilde{U})}, \epsilon_1, \epsilon_1^0)} \big) \Big],$$

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

### Theorem

- **Law invariance.** For any $\xi$ and $\xi'$ $\mathcal{X}$-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 $\mu = \mathbb{P}_{(U,\xi)} \in \mathcal{P}_\lambda(I \times \mathcal{X})$.

### Theorem

- **Law invariance.** For any $\xi$ and $\xi'$ $\mathcal{X}$-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 $V_{\text{weak}}$ fixed point for the operator $\mathcal{T}$:

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

- **Existence of optimal randomized feedback control** $\mathrm{a}^\star$ **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 $\xi$ and $\xi'$ $\mathcal{X}$-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 $V_{\text{weak}}$ fixed point for the operator $\mathcal{T}$:

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

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

$$\alpha_t^\star = \mathrm{a}^\star(\mathbb{P}^0_{(U,X_t)}, U, X_t, \tilde{U}_t) \tag{12}$$

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

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

## Theorem

- **Law invariance.** For any $\xi$ and $\xi'$ $\mathcal{X}$-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 $V_{\text{weak}}$ fixed point for the operator $\mathcal{T}$:

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

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

$$\alpha_t^\star = \mathrm{a}^\star(\mathbb{P}^0_{(U,X_t)}, U, X_t, \tilde{U}_t) \tag{12}$$

  where $(\tilde{U}_t)_{t\in\mathbb{N}}$ sequence of $i.i.d$ uniform random variables for some measurable function $a^*(\mu, u, x, \tilde{u})$ on $\mathcal{P}_\lambda(I \times \mathcal{X}) \times I \times \mathcal{X} \times [0,1]$.
- **Hölder property of the value function.** There exists a positive constant $\gamma \leqslant 1$ such that the value function function is $\gamma$-Hölder ie

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

## Formulation of the strong formulation

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

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

- Value function in the strong formulation :

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

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

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

- 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 $\xi$ be random variables such that $\mathbb{P}_{\xi^u} = \mathbb{P}_{\xi | U = u}$ for $\lambda$ a.e $u \in I$. Then , we have

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

**Proof.**

The main idea of the proof follows from the fact that given an optimal randomized feedback policy $\mathrm{a}$ for the weak formulation $\mathrm{a}$, it gives an optimal feedback control for the strong formulation by setting for the same $\mathrm{a}^\star$ in (12).

$$\alpha_t^{\text{strong},u} = \mathrm{a}^\star(\mathbb{P}^0_{X_t^v}(\mathrm{d}x)\mathrm{d}v, u, X_t^u, \tilde{U}_t^u),$$

since $V_{\text{weak}}^{\alpha^{\text{weak}}}(\xi) = V_{\text{strong}}^{\alpha^{\text{strong}}}(\boldsymbol{\xi})$ when $\alpha^{\text{weak}}$ and $\alpha^{\text{strong}}$ are associated to the same policy $\mathrm{a}$. $\square$

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

# Contents

# 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 $\mathbf{X}^N = (X_i^N)_{i \in [\![1,N]\!]}$

$$\begin{cases} X_0^i = x_0^i, \\ X_{t+1}^i = F_N\Big(\frac{i}{N}, X_t^i, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(\frac{j}{N}, X_t^j, \alpha_t^j)}, \epsilon_{t+1}^j, \epsilon_{t+1}^0\Big), \quad t \in \mathbb{N}. \end{cases} \tag{13}$$

  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.

$$V_N^{\boldsymbol{\alpha}}(\boldsymbol{x}_0) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t f_N\Big(\frac{i}{N}, X_t^i, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(\frac{j}{N}, X_t^j, \alpha_t^j)}\Big)\Big], \tag{14}$$

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

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

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

- State dynamics (13) can be written :

$$\mathbf{X}_{t+1} = \boldsymbol{F}_N(\mathbf{X}_t, \boldsymbol{\alpha}_t, \boldsymbol{\epsilon}_{t+1}), \tag{15}$$

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 (14) for the N agent MDP :

$$V_N^{\alpha}(\boldsymbol{x}_0) = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \beta^t \boldsymbol{f}_N(\mathbf{X}_t, \boldsymbol{\alpha}_t)\Big]. \tag{16}$$

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

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

### Definition of the Bellman operator on $L_m^\infty(\mathcal{X}^N)$

- Bellman operator for the $N$-agent MDP def

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

where

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

### Assumption on the regularity of the initial condition

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

$$d(x^i, x^j) \leqslant C\frac{|i-j|}{N}. \tag{17}$$

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

- The assumption (17) 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

$$I \ni u \mapsto f(u, x, a, \mu) \in \mathbb{R}, \tag{18}$$

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

$$\max_{1 \leqslant j \leqslant N} \sup_{(x,a,\mu) \in \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A)} V_{\frac{j-1}{N}}^{\frac{j}{N}}(f) \leqslant \frac{C}{\sqrt{N}}. \tag{19}$$

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)

$$\mathbb{E}\Big[d\Big(F(u, x, a, \mu, \epsilon_1^1, e^0), F(u', x', a', \mu', \epsilon_1^1, e^0)\Big)\Big] \leqslant K_F\Big(\mathbf{d}((u, x, a), (u', x', a')) + \mathcal{W}(\mu, \mu')\Big), \tag{20}$$

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

**(Convergence of $f_N$ and $F_N$ towards $f$ and $F$).**

There exists two positive decreasing sequences $(\epsilon_N^f)_{N\in\mathbb{N}*}$ and $(\epsilon_N^F)_{N\in\mathbb{N}*}$ converging to 0 as $N\to\infty$ such that

(1)

$$\max_{1\leqslant j\leqslant N}\sup_{(x,a,\mu)\in\mathcal{X}\times A\times\mathcal{P}(I\times\mathcal{X}\times A)}\left|f(\frac{j}{N},x,a,\mu)-f_N(\frac{j}{N},x,a,\mu)\right|\leqslant\epsilon_N^f \tag{21}$$

(2)

$$\max_{1\leqslant j\leqslant N}\sup_{(x,a,\mu)\in\mathcal{X}\times A\times\mathcal{P}(I\times\mathcal{X}\times A)}\mathbb{E}\Big[d\big(F(\frac{j}{N},x,a,\mu,\epsilon_1^i,\epsilon_1^0),F_N(\frac{j}{N},x,a,\mu,\epsilon_1^i,\epsilon_1^0)\big)\Big]\leqslant\epsilon_N^F \tag{22}$$

## 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 function $\widetilde{V}$ defined on $\mathcal{X}^N$ by

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

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

- 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

$$\big|V_N(\boldsymbol{x}) - V(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}])\big| \underset{N \to \infty}{\to} 0. \tag{23}$$

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

$$\big|V_N(\boldsymbol{x}) - V(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}])\big| \leqslant C\Big({M_N}^\gamma + O(N^{-\frac{\gamma}{2}}) + \epsilon_N^f + (\epsilon_N^F)^\gamma\Big).$$

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

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

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

## Theorem : Approximate optimal policies

Let $\mathrm{a}^\star : \mathcal{P}_\lambda(I \times \mathcal{X}) \times I \times \mathcal{X} \times [0,1] \to A$ be an optimal randomized feedback policy for the CNEMF-MDP satisfying the following regularity condition

$$\mathbb{E}\Big[\Big|\Big(\mathfrak{a}(\mu, u, x, U) - \mathfrak{a}_\epsilon(\mu, u', x', U)\Big|\Big] \leqslant K\Big(|u - u'| + |x - x'|\Big), \tag{24}$$

for a positive constant $K \geqslant 0$. Then, denoting

$$\boldsymbol{\pi}_r^{\mathrm{a}^\star, N}(\boldsymbol{x}, \tilde{\boldsymbol{u}}) := \big(\mathrm{a}^\star(\mu_N^\lambda[\boldsymbol{u}, \boldsymbol{x}], \frac{i}{N}, x^i, \tilde{u}^i\big)_{i \in [\![1,N]\!]}, \tag{25}$$

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

$$\boldsymbol{\alpha}_t^{r,N} = \boldsymbol{\pi}_r^{\mathfrak{a}^\star, N}(\mathbf{X}_t, \boldsymbol{U}_t), \quad t \in \mathbb{N}, \tag{26}$$

where $\big\{\boldsymbol{U}_t = (U_t^i)_{i \in \{1,N\}}, t \in \mathbb{N}\big\}$ is a family of mutually i.i.d uniform random variables on $[0,1]$, is an $O(M_N^\gamma + N^{-\frac{\gamma}{2}} + \epsilon_N^f + (\epsilon_N^F)^\gamma)$ 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 $\mathrm{a}^\star$.

## 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 $\mathrm{a}^\star$.
- Convergence of the value function $V_N$ of the $N$-agent MDP towards the value function $V$ for initial state agents $\boldsymbol{x} := (x^i)_{i \in [\![1,N]\!]} \in \mathcal{X}^N_{\text{reg}}$ 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 $\to$ 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 $\mathrm{a}^\star$.
- Convergence of the value function $V_N$ of the $N$-agent MDP towards the value function $V$ for initial state agents $\boldsymbol{x} := (x^i)_{i \in [\![1,N]\!]} \in \mathcal{X}_{\mathrm{reg}}^N$ with explicit rate of convergence
- Optimal randomized feedback control for CNEMF-MDP $\to$ Quantitative approximate optimal policy for the $N$-agent MDP.

## Future works on non exchangeable mean field systems

- Model with controlled interactions $\mathbf{X}^N = (X^{i,N})_{i\in[\![1,N]\!]}$

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

where $\beta^{i,j,N}$ should be represented as the interaction term between agents $i$ and $j$.
$\rightarrow$ Include controlled graphon interactions $\frac{1}{N}\sum_{j=1}^{N} G_N(\frac{i}{N}, \frac{j}{N}, \beta_t^{i,j,N})$.

## Future works on non exchangeable mean field systems

- Model with controlled interactions $\mathbf{X}^N = (X^{i,N})_{i \in [\![1,N]\!]}$

$$\begin{cases} X^{i,N}_{t+1} = F_N\Big(\frac{i}{N}, X^{i,N}_t, \frac{1}{N}\sum_{j=1}^{N} \delta_{(\frac{j}{N}, X^{j,N}_t, \beta^{i,j,N}_t)}, \epsilon^i_{t+1}, \epsilon^0_{t+1}\Big), \quad t \in \mathbb{N}. \\ X^{i,N}_0 = x^i_0, \end{cases}$$

where $\beta^{i,j,N}$ should be represented as the interaction term between agents $i$ and $j$.
$\rightarrow$ Include controlled graphon interactions $\frac{1}{N}\sum_{j=1}^{N} G_N\big(\frac{i}{N}, \frac{j}{N}, \beta^{i,j,N}_t\big)$.

- Numerical algorithms in the context of the label state formulation
  (1) In continuous time :
  $\rightarrow$ In a model based setting : by DPP principle or by backward algorithms based on Pontryagin formulation.
  $\rightarrow$ In a model-free setting : By policy gradient and actor-critic algorithms.

## Future works on non exchangeable mean field systems

- Model with controlled interactions $\mathbf{X}^N = (X^{i,N})_{i \in [\![1,N]\!]}$

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

where $\beta^{i,j,N}$ should be represented as the interaction term between agents $i$ and $j$.
→ Include controlled graphon interactions $\frac{1}{N}\sum_{j=1}^{N} G_N\big(\frac{i}{N}, \frac{j}{N}, \beta_t^{i,j,N}\big)$.

- Numerical algorithms in the context of the label state formulation
  (1) In continuous time :
  → In a model based setting : by DPP principle or by backward algorithms based on Pontryagin formulation.
  → In a model-free setting : By policy gradient and actor-critic algorithms.
  (2) In discrete time :
  → Develop reinforcement learning algorithms based on the DPP principle.

## Future works on non exchangeable mean field systems

- Model with controlled interactions $\mathbf{X}^N = (X^{i,N})_{i\in[\![1,N]\!]}$

$$\begin{cases} X^{i,N}_{t+1} = F_N\Big(\frac{i}{N}, X^{i,N}_t, \frac{1}{N}\sum_{j=1}^{N} \delta_{(\frac{j}{N}, X^{j,N}_t, \beta^{i,j,N}_t)}, \epsilon^i_{t+1}, \epsilon^0_{t+1}\Big), \quad t \in \mathbb{N}. \\ X^{i,N}_0 = x^i_0, \end{cases}$$

  where $\beta^{i,j,N}$ should be represented as the interaction term between agents $i$ and $j$.
  $\rightarrow$ Include controlled graphon interactions $\frac{1}{N}\sum_{j=1}^{N} G_N\big(\frac{i}{N}, \frac{j}{N}, \beta^{i,j,N}_t\big)$.
- Numerical algorithms in the context of the label state formulation
  (1) In continuous time :
  $\rightarrow$ In a model based setting : by DPP principle or by backward algorithms based on Pontryagin formulation.
  $\rightarrow$ In a model-free setting : By policy gradient and actor-critic algorithms.
  (2) In discrete time :
  $\rightarrow$ Develop reinforcement learning algorithms based on the DPP principle.
- LQ control problem for non exchangeable mean field systems (with common noise).

## Problem formulation

In the sequel, we extend a targeted advertising model developped in [3] Section 3.5 to the non exchangeable mean field setting. To this end, let $C$ denotes a company, $I$ an influencer working for the advertising of $C$ which can impact customer choices. We model the state space $\mathcal{X} = \{0, 1\}$ where $x = 1$ (resp $x = 0$) indicates that $x$ is ( is not) a customer of $C$. We also model the action space $A = \{0, 1\}$ where $a = 1$ (resp $a = 0$) indicates if $I$ will display (or not) an ad .

- Reward function : for $u \in I, x \in \mathcal{X}, a \in A$,

$$f(u, x, a) = x - c^u a,$$

  where $c^u$ is an ad cost for agent $u \in I$. It means that if the $u$ labeled user is a customer of C $(x = 1)$, it contributes to the revenue of the company but if $C$ had to send him an ad $(a = 1)$, it costs $c^u$ to the company.

- State transition function. For $(u, x, a) \in I \times \mathcal{X} \times A$, $e \in [0, 1]$ and $\mu \in \mathcal{P}(I \times \mathcal{X})$,

$$F(u, x, a, \mu, e) = \begin{cases} \mathbf{1}_{e > \int_I G(u,v)\mu^v(\{0\})\mathrm{d}v - 2\eta^u a} & \text{if } x = 0, \\ \mathbf{1}_{e < \int_I G(u,v)\mu^v(\{1\})\mathrm{d}v + 2\eta^u a} & \text{if } x = 1. \end{cases}$$

  The parameters of the state transition function $F$ can be interpreted as follows. $\eta^u > 0$ represents the efficiency of an ad to become or remain a customer of $C$ for agent $u$ and a large $e$ indicates an intention to switch operators. The interpretation of the state transition function is then the following: If agent $u$ is in state 0, he's more likely to become a customer of $C$ if $\mu^u(\{0\})$ is low and he receives an ad $(a = 1)$.

# 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