Learning Mean-Field Games

Berkay Anahtarcı

Özyeğin University, Department of Natural and Mathematical Sciences

Based on joint work with
Can Deha Karıksız and Naci Saldi

January 23, 2024

Outline

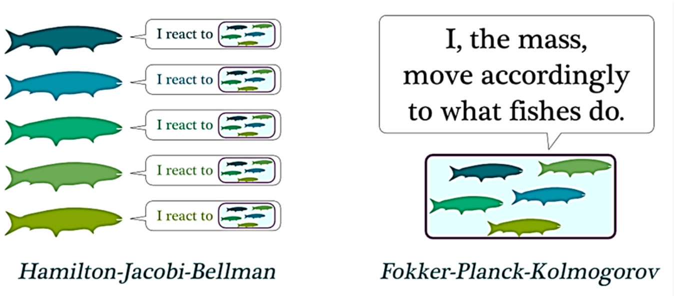
1. Motivating Example
2. N-Player Games
3. Mean-Field Games
4. Value Iteration Algorithm
5. Reinforcement Learning for MFGs

Overview of Mean-Field Game (MFG)

Mean-Field Games (MFGs) are characterized as

- A game involving a vast population of small interacting individuals:
  - **Large population**: Encompassing a continuum of players.
  - **Small interacting**: Strategies based on aggregated macroscopic information (mean-field).
- Originated from the study of weakly interacting particles in physics.
- Theoretical groundwork laid by Huang, Malhamé, and Caines (2006), and Lasry and Lions (2007).
- **Main idea:** In an N-player game, as N grows, the "aggregated" version, MFG, approximates the game using the *Law of Large Numbers*, in terms of ϵ-Nash equilibrium.

Motivating Example: Crowd Motion Analysis



https://www.science4all.org/article/mean-field-games/

Classical N-Player Markovian Games

- Given the current **state profile** of N-players $\boldsymbol{x}_t = (x_t^1, \ldots, x_t^N) \in \mathcal{X}^N$ and the action $a_t^i \in \mathcal{A}$, player i receives a **reward** $r^i(\boldsymbol{x}_t, a_t^i)$.
- Their state changes to x_{t+1}^i according to a **transition probability** function $P^i(\boldsymbol{x}_t, a_t^i)$.
- The **policy** $\pi_t^i : \mathcal{X}^N \to \Delta_{\mathcal{A}}$ maps each state profile $\boldsymbol{x} \in \mathcal{X}^N$ to a randomized action, with $\Delta_{\mathcal{A}}$ the space of probability measures on space $\mathcal{A}$.
- In a Markovian game, the admissible policy/control for player i is determined by the current state: $a_t^i = \pi_t^i(\boldsymbol{x}_t)$.

Classical N-Player Markovian Games

Problem Formulation

$$\text{maximize}_{\boldsymbol{\pi}} \quad V^i(\boldsymbol{x}, \boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(\boldsymbol{x}_t, a_t^i) \mid \boldsymbol{x}_0 = \boldsymbol{x}\right]$$

$$\text{subject to} \quad x_{t+1}^i \sim P^i(\boldsymbol{x}_t, a_t^i) \text{ and } a_t^i \sim \pi_t^i(\boldsymbol{x}_t)$$

- $V^i(\boldsymbol{x}, \boldsymbol{\pi})$ is the **value function** for player i, given the initial state profile $\boldsymbol{x}$ and the **policy profile** sequence $\boldsymbol{\pi} := \{\boldsymbol{\pi}_t\}_{t=0}^{\infty}$ with $\boldsymbol{\pi}_t = (\pi_t^1, \ldots, \pi_t^N)$.
- $\gamma \in (0, 1)$ is the **discount factor**.

N-Player Games

Definition (N-Player Game: Nash Equilibrium)

Nash Equilibrium (NE) consists of strategies where no agent can gain an advantage from unilaterally deviating from this set of strategies. Formally, $\boldsymbol{\pi}^*$ is an NE if for all i and $\boldsymbol{x}$,

$$V^i(\boldsymbol{x}, \boldsymbol{\pi}^*) \geq V^i(\boldsymbol{x}, (\pi_1^*, \ldots, \pi_i, \ldots, \pi_N^*))$$

holds for any $\pi_i : \mathcal{X}^N \to \Delta_{\mathcal{A}}$.

From N-Player Game to MFG

- Assume all players are identical, indistinguishable and interchangeable.
- Each player has a negligible impact on the rest of the population.
- One can view the limit of other players' states $\boldsymbol{x}_t^{-i} := (x_t^1, \ldots, x_t^{i-1}, x_t^{i+1}, \ldots, x_t^N)$ as a population state distribution

$$\mu_t(x) := \lim_{N\to\infty} \frac{\sum_{j=1, j\neq i}^{N} \mathbb{1}_{x_t^j = x}}{N}.$$

- Due to the homogeneity of the players, one can then focus on a single (representative) player.

From N-Player Game to MFG

MFG Formulation

$$\text{maximize}_{\boldsymbol{\pi}} \quad V(x, \boldsymbol{\pi}, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t, \mu_t) \mid x_0 = x\right]$$

$$\text{subject to} \quad x_{t+1} \sim P(x_t, a_t, \mu_t) \text{ and } a_t \sim \pi_t(x_t, \mu_t)$$

- Here $\boldsymbol{\pi} := \{\pi_t\}_{t=0}^{\infty}$ denotes the **policy sequence** and $\boldsymbol{\mu} := \{\mu_t\}_{t=0}^{\infty}$ the **distribution flow**.
- In the MFG setting, at time t, after the representative player chooses their action according to some policy π_t, they will receive reward $r(x_t, a_t, \mu_t)$ and their state will evolve under $P(\cdot \mid s_t, a_t, \mu_t)$.
- Here $\pi : \mathcal{X} \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{A}}$.

Population Dynamics

Definition (McKean–Vlasov Equation)

The evolution of the population is given by a transition matrix defined by

$$\mu_{t+1}(y) = \sum_{x \in \mathcal{X}} \mu_t(x) \sum_{a \in \mathcal{A}} \pi_t(a \mid x) p(y \mid x, a, \mu_t) := P_t^{\pi} \mu_t(y)$$

for all $\pi_t \in \Pi$, $\mu_t \in \Delta_{\mathcal{X}}$ and $x \in \mathcal{X}$.

Stationary MFGs

- Assume that the players interact through a **stationary** distribution, which represents a steady state of the population.
- The model is defined by a tuple $(\mathcal{X}, \mathcal{A}, p, r, \gamma)$ consisting of
  - a state space $\mathcal{X}$ and an action space $\mathcal{A}$,
  - a one-step transition probability kernel $p : \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{X}}$,
  - a one-step reward function $r : \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \to \mathbb{R}$,
  - and a discount factor $\gamma \in [0, 1]$.
- The state of the population is given by $\mu_t = \mu \in \Delta_{\mathcal{X}}$ for all t.
- Consider a representative agent using policy $\pi \in \Pi$.

Stationary MFGs

Definition (Total discounted reward)

$$J(\pi, \mu) = \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n r(x_n, a_n, \mu)\right]$$

$$x_0 \sim \mu, \quad x_{n+1} \sim p(\cdot \mid x_n, a_n, \mu), \quad a_n \sim \pi(\cdot \mid x_n).$$

- Given a population state, the goal for a representative agent, is to find the best reaction, i.e., a policy that maximizes their total reward.

Stationary MFGs

Definition (Best Response Map)

$$\Psi : \Delta_{\mathcal{X}} \to 2^{\Pi}, \quad \mu \mapsto \Psi(\mu) := \operatorname*{argmax}_{\pi \in \Pi} J(\pi, \mu) \subseteq \Pi.$$

Definition (Population Behaviour Map)

$$\Lambda : \Pi \to 2^{\Delta_{\mathcal{X}}}, \quad \pi \mapsto \Lambda(\pi) := \{\mu \in \Delta_{\mathcal{X}} \mid \mu = P^{\pi}\mu\}$$

is the **stationary distribution** obtained when using π (that we assume to be unique).

Stationary MFGs

Definition (Stationary MF Nash Equilibrium)

A pair $(\pi_*, \mu_*) \in \Pi \times \Delta_{\mathcal{X}}$ is called **stationary MFNE** if it satisfies:

$$\pi_* \in \Psi(\mu_*) \quad \text{and} \quad \mu_* \in \Lambda(\pi_*)$$

Alternatively, an equilibrium can be defined as a fixed point:

- π_* is a **stationary MFNE policy** if it is a fixed point of $\Psi \circ \Lambda$,
- μ_* is a **stationary MFNE distribution** if it is the stationary distribution of a stationary MFNE policy.

Stationary MFGs

Definition (State-Action Value Function)

The state-action value function associated to a stationary policy π for a given distribution μ is defined as:

$$Q^{\pi,\mu}(x, a) = \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n r(x_n, a_n, \mu) \mid x_0 = x, a_0 = a\right]$$

where $x_{n+1} \sim p(\cdot \mid x_n, a_n, \mu)$ and $a_n \sim \pi(\cdot \mid x_n)$.

- $Q^{\pi,\mu}$ satisfies the fixed point equation: $Q = B^{\pi,\mu} Q$.

Stationary MFGs

Definition (Bellman Operator)

$$(B^{\pi,\mu}Q)(x,a) = r(x,a,\mu) + \gamma \sum_{x'} p(x' \mid x,a,\mu) \sum_{a'} \pi(a' \mid x') Q(x',a')$$

- Note that

$$\sum_{x'} p(x' \mid x,a,\mu) \sum_{a'} \pi(a' \mid x') Q(x',a') = \mathbb{E}_{\substack{x' \sim p(\cdot|x,a,\mu) \\ a' \sim \pi(\cdot|x')}} [Q(x',a')].$$

Stationary MFGs

Definition (Optimal State-Action Value Function)

$$Q^{*,\mu}(x, a) = \sup_{\pi} Q^{\pi,\mu}(x, a)$$

- It satisfies the fixed point equation: $Q = B^{*,\mu} Q$.

Optimal Bellman Operator associated to μ

$$(B^{*,\mu} Q)(x, a) = r(x, a, \mu) + \gamma \mathop{\mathbb{E}}_{x' \sim p(\cdot|x,a,\mu)} [\max_{a'} Q(x', a')]$$

- Here

$$\mathop{\mathbb{E}}_{x' \sim p(\cdot|x,a,\mu)} [\max_{a'} Q(x', a')] = \sum_{x'} p(x' \mid x, a, \mu) \max_{a'} Q(x', a').$$

Solving MFGs

Best Response-Based Methods

Let μ_0 be given, for $i = 0, \ldots, L-1$:

$$\begin{cases} \pi_{i+1} = \Psi(\mu_i) \\ \mu_{i+1} = \Pi(\pi_{i+1}) \end{cases}$$

Under suitable conditions, (π_L, μ_L) is close to (π_*, μ_*) when L is large enough.

Transition Matrix Approximation

Let μ_0 be given, for $i = 0, \ldots, L-1$:

$$\begin{cases} \pi_{i+1} = \Psi(\mu_i) \\ \mu_{i+1} = P^{\pi^{i+1}} \mu_i \end{cases}$$

Value Iteration Algorithm [A., Karıksız, Saldi (2021)]

Cost Function

Given μ, the cost of policy π with initial state x is:

$$J_\mu(\pi, x) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \beta^t c(x(t), a(t), \mu) \mid x(0) = x \right]$$

Bellman Optimality Operator

$$J_\mu^*(x) = \min_a \left[c(x, a, \mu) + \beta \sum_y^{\infty} J_\mu^*(y) p(y \mid x, a, \mu) \right]$$

- The optimal cost is given by $J_\mu^* = \inf_\pi J_\mu(\pi, x)$.
- J_μ^* is the unique fixed point of the Bellman optimality operator which is β-contractive.
- If $\pi_\mu : \mathcal{X} \to \mathcal{A}$ attains the minimum, then it is optimal.

Value Iteration Algorithm

- We can also characterize π_μ using Q-functions.

Optimal Q-function

$$Q_\mu^*(x, a) = c(x, a, \mu) + \beta \sum_y^\infty J_\mu^*(y) p(y \mid x, a, \mu)$$

- Then $Q_{\mu,\min}^*(x) := \min_a Q_\mu^*(x, a) = J_\mu^*(x)$.
- $Q_\mu^*(x, a)$ is the unique fixed point of the β-contractive operator:

$$Q_\mu^*(x, a) = c(x, a, \mu) + \beta \sum_y^\infty Q_{\mu,\min}^*(x) p(y \mid x, a, \mu)$$

- If $\pi_\mu(x) = \operatorname{argmin}_a Q_\mu^*(x, a)$. Then π_μ is optimal.

Value Iteration Algorithm

Optimal Q-function for μ

$$H_1 : \mu \to Q^*_\mu$$

New mean-field

$$H_2 : (\mu, Q) \mapsto \sum_x p(\cdot \mid x, \pi_Q(x), \mu)\mu(x)$$

$$\pi_Q(x) := \operatorname*{argmin}_a Q(x, a) \quad \text{[greedy policy]}$$

Mean-Field Equilibrium (MFE)

$$H : \mu \mapsto H_2(\mu, H_1(\mu)) = \sum_x p(\cdot \mid x, \pi_\mu(x), \mu)\mu(x)$$

Value Iteration Algorithm

- It turns out that H is a contraction.
- Using the Banach Fixed Point theorem, the VI algorithm gives the fixed point μ_* and the corresponding

VI Algorithm

Start with μ_0
while $\mu_n \neq \mu_{n-1}$ **do**
  $\mu_{n+1} = H(\mu_n)$
end while
return Fixed-point μ_* of H and $Q^*_{\mu_*} = H_1(\mu_*)$

- If $(\mu_*, Q^*_{\mu_*})$ is the output of the value iteration algorithm above, then the pair (μ_*, π_{μ_*}) is a mean-field equilibrium.

Value Iteration Algorithm

Assumptions

- The one-stage cost c function and the transition kernel p are Lipschitz continuous.
- $F(x, \nu, \mu, \cdot) := c(x, \cdot, \mu) + \beta \sum_{y \in \mathcal{X}} \nu(y) p(y \mid x, \cdot, \mu)$ is ρ-strongly convex. Moreover, its gradient $\nabla F(x, \nu, \mu, \cdot)$ with respect to a is Lipschitz continuous.

Learning Algorithm [A., Karıksız, Saldi (2021)]

- If p and c are unknown, one needs to develop a learning algorithm to compute a mean-field equilibrium.
- When the model is known, given μ, the MFE operator H is composition of H_1 and H_2:
  - $H_1(\mu)$ is the optimal Q-function Q_μ^* for μ
  - $H_2(\mu, Q_\mu^*)$ is the new mean-field term.
- When the model is unknown, we replace H_1 and H_2 with random operators $\hat{H}_1$ and $\hat{H}_2$.

Algorithm for $\hat{H}_1$

Fitted Q-learning

Inputs $([N, L], \mu)$
Generate i.i.d. samples $\{(x_t, a_t)\}_{t=1}^{N}$ and let
$c_t = c(x_t, a_t, \mu),\ y_{t+1} \sim p(\cdot \mid x_t, a_t, \mu)$.
Start with $Q_0 = 0$
for $i = 0, \ldots, L-1$ *do*

$$Q_{i+1} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{t=1}^{N} \left(f(x_t, a_t) - c_t + \beta \min_{a'} Q_i(y_{t+1}, a') \right)^2 \right]$$

end for
return Q_L

Algorithm for $\hat{H}_2$

Simulation

Inputs (M, μ, Q)

for $x \in \mathcal{X}$ **do**

Generate i.i.d. samples $\{y_t^x\}_{t=1}^M$ using $y_t^x \sim p(\cdot \mid x, \pi_Q(x), \mu)$ and define

$$p_M(\cdot \mid x, \pi_Q(x), \mu) = \frac{1}{M} \sum_{t=1}^{M} \delta_{y_t^x}(\cdot)$$

end for

return $\sum_x p_M(\cdot \mid x, \pi_Q(x), \mu)\mu(x)$

Approximate MFE operator $\hat{H}$

Learning Algorithm

Inputs $(K, \{[N_k, L_k]\}_{k=0}^{K}, \{M_k\}_{k=0}^{K}, \mu_0)$
Start with μ_0
for $k = 0, \ldots K-1$ **do**

$$\mu_{k+1} = \hat{H}([N_k, L_k], M_k)(\mu_k) := \hat{H}_2[M_k](\mu_k, \hat{H}_1[N_k, L_k](\mu_k))$$

end for
return μ_k and $Q_k = \hat{H}_1([N_k, L_k])(\mu_k)$

Main Results

Approximate Mean-Field Equilibrium

Let (μ_k, Q_k) be the output of the learning algorithm $\hat{H}$. Define $\pi_K(x) := \text{argmin}_a Q_K(x, a)$. Then, with probability at least $1 - \delta$,

$$\sup_x \|\pi_K(x) - \pi_*(x)\| \leq \kappa(\epsilon, \Delta)$$

where $\kappa(\epsilon, \Delta) = O(\epsilon + \Delta)$.

Approximate Nash Equilibrium

Let π_K be the policy obtained from the learning algorithm. Then, for any $\delta > 0$, there exists a positive integer $N(\delta)$ such that for each $N \geq N(\delta)$, the N-tuple of policies $\boldsymbol{\pi}^{(N)} = \{\pi_K, \pi_K, \ldots, \pi_K\}$ is an $(\delta + \tau\kappa(\epsilon, \Delta))$-Nash equilibrium for the game with N agents, with probability at least $1 - \delta$.

References

Berkay Anahtarci, Can Deha Karıksız, Naci Saldi
Learning Mean-Field Games with Discounted and Average-Costs
Journal of Machine Learning Research **24**(17):1-59, 2023.

Mathieu Laurière, Sarah Perrin, Matthieu Geisty, Olivier Pietquiny
Learning Mean Field Games: A Survey
arXiv: 2205.12944v1 (2022)

Xin Guo, Anran Hu, Renyuan Xu, Junzi Zhang
Learning Mean-Field Games
arXiv: 1901.09585v4 (2021)

The End