Mean Field Game and Mean Field Control Q-Learning

Jean-Pierre Fouque

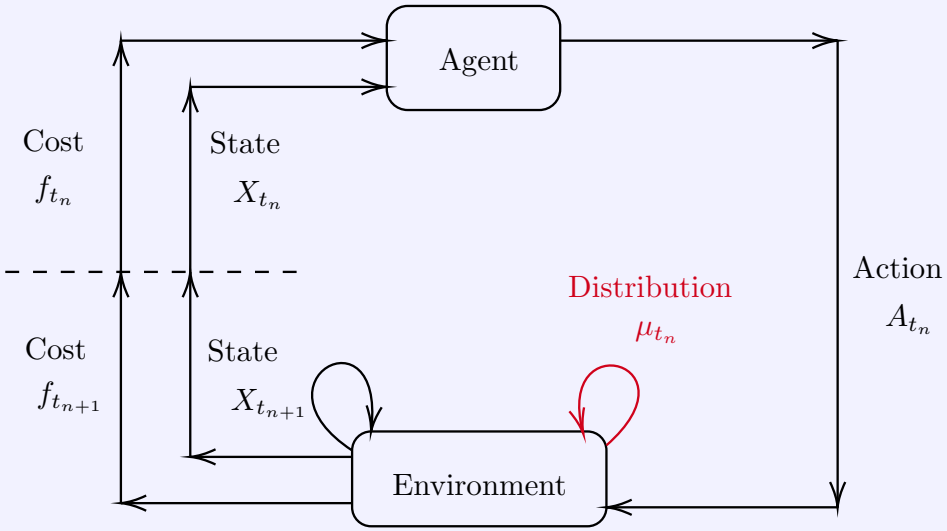
Department of Statistics and Applied Probability
University of California, Santa Barbara

Collaborators:
Andrea Angiuli, Mathieu Laurière, and Mengrui Zhang
https://arxiv.org/abs/2312.06659

IMSI
May 18-22, 2026

Reinforcement Learning Setup

- Infinite Horizon Discrete Time
- Finite State Space $\mathcal{X}$ and Finite Action Space $\mathcal{A}$
- Reward (or Cost) Function $f(x, a, \mu)$ which depends on the **population distribution** μ
- Transition probabilities $p(\cdot \mid x, a, \mu)$
- Q-matrix $Q(x, a)$ updated by the agent



Policies and Q-functions

Let $\pi : \mathcal{X} \to \Delta^{|\mathcal{A}|}$ be a policy taking the state x as input and outputting a probability distribution over the action set. Given a population distribution μ, the Q-function for a single agent is

$$Q_\mu^\pi(x,a) := f(x,a,\mu) + \mathbb{E}\left[\sum_{n\geq 1} \gamma^n f(X_n^{\pi,\mu}, A_n, \mu)\right], \quad \gamma < 1$$

where

$$\begin{cases} X_0^{\pi,\mu} = x, A_0 = a, \\ A_n \sim \pi(\cdot|X_n^{\pi,\mu}), \qquad n \geq 1 \\ X_{n+1}^{\pi,\mu} \sim p(\cdot|X_n^{\pi,\mu}, A_n, \mu), \qquad n \geq 0 \end{cases}$$

We restrict to Π the set of all admissible policies which are such that for any μ the controlled process $(X_n^{\pi,\mu})$ has a limiting distribution, i.e. $\lim_{n\to\infty} \mathcal{L}(X_n^{\pi,\mu})$ exists. For a finite state Markov chain this is the case if $(X_n^{\pi,\mu})$ is irreducible and aperiodic under proper assumptions on the transition kernel p.

Policies and Q-functions

Let $\pi : \mathcal{X} \to \Delta^{|\mathcal{A}|}$ be a policy taking the state x as input and outputting a probability distribution over the action set. Given a population distribution μ, the Q-function for a single agent is

$$Q_\mu^\pi(x,a) := f(x,a,\mu) + \mathbb{E}\left[\sum_{n\geq 1} \gamma^n f(X_n^{\pi,\mu}, A_n, \mu)\right], \quad \gamma < 1$$

where

$$\begin{cases} X_0^{\pi,\mu} = x, A_0 = a, \\ A_n \sim \pi(\cdot|X_n^{\pi,\mu}), \qquad n \geq 1 \\ X_{n+1}^{\pi,\mu} \sim p(\cdot|X_n^{\pi,\mu}, A_n, \mu), \qquad n \geq 0 \end{cases}$$

We restrict to Π the set of all admissible policies which are such that for any μ the controlled process $(X_n^{\pi,\mu})$ has a limiting distribution, i.e. $\lim_{n\to\infty} \mathcal{L}(X_n^{\pi,\mu})$ exists. For a finite state Markov chain this is the case if $(X_n^{\pi,\mu})$ is irreducible and aperiodic under proper assumptions on the transition kernel p.

Mean Field Game Formulation

For fixed μ, the optimal Q-function is defined as:

$$Q_\mu^*(x,a) := \inf_{\pi \in \Pi} Q_\mu^\pi(x,a)$$

which satisfies the Bellman equation:

$$Q_\mu^*(x,a) = f(x,a,\mu) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\mu) \min_{a'} Q_\mu^*(x',a')$$

An optimal policy satisfies

$$Q_\mu^{\pi^*}(x,a) = Q_\mu^*(x,a), \qquad \text{for all } (x,a).$$

Definition

$(\hat{\mu}, \hat{\pi})$ forms a **Nash equilibrium** if the following two statements hold:

1. (best response) $Q_{\hat{\mu}}^{\hat{\pi}} = Q_{\hat{\mu}}^*$
2. (consistency) $\hat{\mu} = \lim_{n \to \infty} \mathcal{L}(X_n^{\hat{\pi}, \hat{\mu}})$.

Mean Field Game Formulation

For fixed μ, the optimal Q-function is defined as:

$$Q_\mu^*(x,a) := \inf_{\pi \in \Pi} Q_\mu^\pi(x,a)$$

which satisfies the Bellman equation:

$$Q_\mu^*(x,a) = f(x,a,\mu) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\mu) \min_{a'} Q_\mu^*(x',a')$$

An optimal policy satisfies

$$Q_\mu^{\pi^*}(x,a) = Q_\mu^*(x,a), \qquad \text{for all } (x,a).$$

Definition

$(\hat{\mu}, \hat{\pi})$ forms a **Nash equilibrium** if the following two statements hold:

1. (best response) $Q_{\hat{\mu}}^{\hat{\pi}} = Q_{\hat{\mu}}^*$
2. (consistency) $\hat{\mu} = \lim_{n \to \infty} \mathcal{L}(X_n^{\hat{\pi}, \hat{\mu}})$.

Two-timescale Mean Field Game Q-learning

Require: N_{steps}: number of steps; ϕ: parameter for soft-min policy; $(\rho^Q_{n,x,a})_{n,x,a}$: learning rates for the value function; $(\rho^\mu_{n,x,a})_{n,x,a}$: learning rates for the population distribution:
MFG: $\rho^\mu_{n,X_n,A_n} \ll \rho^Q_{n,X_n,A_n}$

1: **Initialization**: $Q_0(x,a) = 0$ and $\mu_0 = [\frac{1}{|\mathcal{X}|}, ..., \frac{1}{|\mathcal{X}|}]$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.
2: **Observe** $X_0 \sim \mu_0$
3: **Update mean field:** $\mu_0 = \boldsymbol{\delta}(X_0)$
4: **for** $n = 0, 1, 2, \ldots, N_{steps} - 1$ **do**
5: **Choose action** $A_n \sim \text{soft-min}_\phi\, Q_n(X_n, \cdot)$ and observe the new state: $X_{n+1} \sim p(\cdot|X_n, A_n, \mu_n)$ provided by the environment
6: **Update population distributions:** $\mu_{n+1} = \mu_n + \rho^\mu_{n,X_n,A_n}(\boldsymbol{\delta}(X_{n+1}) - \mu_n)$
7: **Observe** the cost $f_{n+1} = f(X_n, A_n, \mu_n)$
8: **Update value function:** $Q_{n+1}(x,a) = Q_n(x,a)$ for all $(x,a) \neq (X_n, A_n)$, and

$$Q_{n+1}(X_n, A_n) = Q_n(X_n, A_n) + \rho^Q_{n,X_n,A_n}\left[f_{n+1} + \gamma \min_{a' \in \mathcal{A}} Q_n(X_{n+1}, a') - Q_n(X_n, A_n)\right]$$

9: **end for**
10: **Return** $(\mu_{N_{steps}}, Q_{N_{steps}})$

Key features: stochastic approximation along one long trajectory and asynchronous

Two-timescale Mean Field Game Q-learning

Require: N_{steps}: number of steps; ϕ: parameter for soft-min policy; $(\rho^Q_{n,x,a})_{n,x,a}$: learning rates for the value function; $(\rho^\mu_{n,x,a})_{n,x,a}$: learning rates for the population distribution:
MFG: $\rho^\mu_{n,X_n,A_n} \ll \rho^Q_{n,X_n,A_n}$

1: **Initialization**: $Q_0(x,a) = 0$ and $\mu_0 = [\frac{1}{|\mathcal{X}|}, ..., \frac{1}{|\mathcal{X}|}]$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.
2: **Observe** $X_0 \sim \mu_0$
3: **Update mean field:** $\mu_0 = \boldsymbol{\delta}(X_0)$
4: **for** $n = 0, 1, 2, \ldots, N_{steps} - 1$ **do**
5: **Choose action** $A_n \sim \text{soft-min}_\phi\, Q_n(X_n, \cdot)$ and observe the new state: $X_{n+1} \sim p(\cdot|X_n, A_n, \mu_n)$ provided by the environment
6: **Update population distributions:** $\mu_{n+1} = \mu_n + \rho^\mu_{n,X_n,A_n}(\boldsymbol{\delta}(X_{n+1}) - \mu_n)$
7: **Observe** the cost $f_{n+1} = f(X_n, A_n, \mu_n)$
8: **Update value function:** $Q_{n+1}(x,a) = Q_n(x,a)$ for all $(x,a) \neq (X_n, A_n)$, and

$$Q_{n+1}(X_n, A_n) = Q_n(X_n, A_n) + \rho^Q_{n,X_n,A_n}\left[f_{n+1} + \gamma \min_{a' \in \mathcal{A}} Q_n(X_{n+1}, a') - Q_n(X_n, A_n)\right]$$

9: **end for**
10: **Return** $(\mu_{N_{steps}}, Q_{N_{steps}})$

Key features: stochastic approximation along one long trajectory and asynchronous

Idealized Two-timescale Iteration Approach

We consider the following iterative procedure, where both variables are updated at each iteration but with different rates, denoted by ρ_n^μ and ρ_n^Q. Starting from an initial guess (Q_0, μ_0), define iteratively for $n = 0, 1, \ldots$:

$$\begin{cases} \mu_{n+1} = \mu_n + \rho_n^\mu \mathcal{P}_2(Q_n, \mu_n) \\ Q_{n+1} = Q_n + \rho_n^Q \mathcal{T}_2(Q_n, \mu_n) \end{cases}$$

where $\mathcal{P}_2 : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \Delta^{|\mathcal{X}|}$ and $\mathcal{T}_2 : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$ are defined by:

$$\begin{cases} \mathcal{P}_2(Q,\mu)(x) = \sum_{x'} \mu(x') p(x|x', \text{soft-min}_\phi\, Q(x'), \mu) - \mu(x) \\ \mathcal{T}_2(Q,\mu)(x,a) = f(x,a,\mu) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') - Q(x,a) \end{cases}$$

Using the notation $(\tilde{\mu}\mathrm{P}^{\pi,\mu})(x) = \sum_{x'\in\mathcal{X}} \tilde{\mu}(x') \sum_{a\in\mathcal{A}} \pi(a|x') p(x|x',a,\mu)$,

$$\begin{cases} \mathcal{P}_2(Q,\mu) = \mu \mathrm{P}^{\text{soft-min}_\phi\, Q,\mu} - \mu \\ \mathcal{T}_2(Q,\mu) = \mathcal{B}_\mu Q - Q, \quad (\mathcal{B} \text{ for Bellman}) \end{cases}$$

Idealized Two-timescale Iteration Approach

We consider the following iterative procedure, where both variables are updated at each iteration but with different rates, denoted by ρ_n^μ and ρ_n^Q. Starting from an initial guess (Q_0, μ_0), define iteratively for $n = 0, 1, \ldots$:

$$\begin{cases} \mu_{n+1} = \mu_n + \rho_n^\mu \mathcal{P}_2(Q_n, \mu_n) \\ Q_{n+1} = Q_n + \rho_n^Q \mathcal{T}_2(Q_n, \mu_n) \end{cases}$$

where $\mathcal{P}_2 : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \Delta^{|\mathcal{X}|}$ and $\mathcal{T}_2 : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$ are defined by:

$$\begin{cases} \mathcal{P}_2(Q,\mu)(x) = \sum_{x'} \mu(x') p(x|x', \text{soft-min}_\phi\, Q(x'), \mu) - \mu(x) \\ \mathcal{T}_2(Q,\mu)(x,a) = f(x,a,\mu) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') - Q(x,a) \end{cases}$$

Using the notation $(\tilde{\mu} \mathrm{P}^{\pi,\mu})(x) = \sum_{x' \in \mathcal{X}} \tilde{\mu}(x') \sum_{a \in \mathcal{A}} \pi(a|x') p(x|x', a, \mu)$,

$$\begin{cases} \mathcal{P}_2(Q,\mu) = \mu \mathrm{P}^{\text{soft-min}_\phi\, Q, \mu} - \mu \\ \mathcal{T}_2(Q,\mu) = \mathcal{B}_\mu Q - Q, \quad (\mathcal{B} \text{ for Bellman}) \end{cases}$$

Multiscale Borkar's Approach

Our proof requires the functions $\mathcal{T}_2$ and $\mathcal{P}_2$ to be Lipschitz continuous. Accordingly, we use soft-min_ϕ instead of $\arg\min$ in $\mathcal{P}_2$ and we require Lipschitz continuity of f.

Following Borkar, if $\rho_n^\mu/\rho_n^Q \to 0$ as $n \to +\infty$, the solution (μ_n, Q_n) to the difference equations has a behavior that is described by the following system of ODEs:

$$\begin{cases} \dot{\mu}_t = \mathcal{P}_2(Q_t, \mu_t) \\ \dot{Q}_t = \frac{1}{\epsilon}\mathcal{T}_2(Q_t, \mu_t) \end{cases}$$

in the regime where

$$\rho_n^\mu/\rho_n^Q \sim \epsilon \ll 1 \quad \text{and} \quad t \to \infty$$

Multiscale Borkar's Approach

Our proof requires the functions $\mathcal{T}_2$ and $\mathcal{P}_2$ to be Lipschitz continuous. Accordingly, we use soft-min_ϕ instead of $\arg\min$ in $\mathcal{P}_2$ and we require Lipschitz continuity of f.
Following Borkar, if $\rho_n^\mu / \rho_n^Q \to 0$ as $n \to +\infty$, the solution (μ_n, Q_n) to the difference equations has a behavior that is described by the following system of ODEs:

$$\begin{cases} \dot{\mu}_t = \mathcal{P}_2(Q_t, \mu_t) \\ \dot{Q}_t = \frac{1}{\epsilon}\mathcal{T}_2(Q_t, \mu_t) \end{cases}$$

in the regime where

$$\rho_n^\mu / \rho_n^Q \sim \epsilon \ll 1 \quad \text{and} \quad t \to \infty$$

Convergence of the Idealized Two-timescale Iteration Scheme

Lipschitz assumptions:

$$|f(x,a,\mu) - f(x,a,\mu')| \leq L_f \|\mu - \mu'\|_1,$$
$$\|p(\cdot|x,a,\mu) - p(\cdot|x,a,\mu')\|_1 \leq L_p \|\mu - \mu'\|_1,$$
$$\|p(\cdot|x,\pi(x),\mu) - p(\cdot|x,\pi(x),\mu')\|_1 \leq \lambda \|\mu - \mu'\|_1.$$

Dobrushin condition: there exists $\alpha > 0$ such that for all $x \in \mathcal{X}$ and $\pi \in \Pi$

$$\sup_{\mu,\mu'} \|p(\cdot|x,\pi(x),\mu) - p(\cdot|y,\pi(y),\mu')\|_1 \leq 2(1-\alpha).$$

By Butkovsky (2014), an optimal condition for the existence of an invariant distribution is $\lambda \in [0, \alpha)$.

Under these assumptions, we can show that the functions $\mathcal{P}_2$ and $\mathcal{T}_2$ are Lipschitz continuous as follows:

Convergence of the Idealized Two-timescale Iteration Scheme

Lipschitz assumptions:

$$
\begin{aligned}
|f(x,a,\mu) - f(x,a,\mu')| &\leq L_f \|\mu - \mu'\|_1, \\
\|p(\cdot|x,a,\mu) - p(\cdot|x,a,\mu')\|_1 &\leq L_p \|\mu - \mu'\|_1, \\
\|p(\cdot|x,\pi(x),\mu) - p(\cdot|x,\pi(x),\mu')\|_1 &\leq \lambda \|\mu - \mu'\|_1.
\end{aligned}
$$

Dobrushin condition: there exists $\alpha > 0$ such that for all $x \in \mathcal{X}$ and $\pi \in \Pi$

$$
\sup_{\mu,\mu'} \|p(\cdot|x,\pi(x),\mu) - p(\cdot|y,\pi(y),\mu')\|_1 \leq 2(1-\alpha).
$$

By Butkovsky (2014), an optimal condition for the existence of an invariant distribution is $\lambda \in [0, \alpha)$.
Under these assumptions, we can show that the functions $\mathcal{P}_2$ and $\mathcal{T}_2$ are Lipschitz continuous as follows:

Convergence of the Idealized Two-timescale Iteration Scheme

Lipschitz assumptions:

$$
\begin{aligned}
|f(x,a,\mu) - f(x,a,\mu')| &\leq L_f \|\mu - \mu'\|_1, \\
\|p(\cdot|x,a,\mu) - p(\cdot|x,a,\mu')\|_1 &\leq L_p \|\mu - \mu'\|_1, \\
\|p(\cdot|x,\pi(x),\mu) - p(\cdot|x,\pi(x),\mu')\|_1 &\leq \lambda \|\mu - \mu'\|_1.
\end{aligned}
$$

Dobrushin condition: there exists $\alpha > 0$ such that for all $x \in \mathcal{X}$ and $\pi \in \Pi$

$$
\sup_{\mu,\mu'} \|p(\cdot|x,\pi(x),\mu) - p(\cdot|y,\pi(y),\mu')\|_1 \leq 2(1-\alpha).
$$

By Butkovsky (2014), an optimal condition for the existence of an invariant distribution is $\lambda \in [0, \alpha)$.

Under these assumptions, we can show that the functions $\mathcal{P}_2$ and $\mathcal{T}_2$ are Lipschitz continuous as follows:

Convergence of the Idealized Two-timescale Iteration Scheme

Lipschitz assumptions:

$$|f(x,a,\mu) - f(x,a,\mu')| \leq L_f \|\mu - \mu'\|_1,$$
$$\|p(\cdot|x,a,\mu) - p(\cdot|x,a,\mu')\|_1 \leq L_p \|\mu - \mu'\|_1,$$
$$\|p(\cdot|x,\pi(x),\mu) - p(\cdot|x,\pi(x),\mu')\|_1 \leq \lambda \|\mu - \mu'\|_1.$$

Dobrushin condition: there exists $\alpha > 0$ such that for all $x \in \mathcal{X}$ and $\pi \in \Pi$

$$\sup_{\mu,\mu'} \|p(\cdot|x,\pi(x),\mu) - p(\cdot|y,\pi(y),\mu')\|_1 \leq 2(1-\alpha).$$

By Butkovsky (2014), an optimal condition for the existence of an invariant distribution is $\lambda \in [0, \alpha)$.
Under these assumptions, we can show that the functions $\mathcal{P}_2$ and $\mathcal{T}_2$ are Lipschitz continuous as follows:

Lipschitz Properties of $\mathcal{P}_2$ and $\mathcal{T}_2$

The function $(Q, \mu) \mapsto \mathcal{P}_2(Q, \mu)$ is Lipschitz with respect to both Q and μ:

$$\|\mathcal{P}_2(Q, \mu) - \mathcal{P}_2(Q', \mu)\|_\infty \leq \phi|\mathcal{A}|\|Q - Q'\|_\infty$$
$$\|\mathcal{P}_2(Q, \mu) - \mathcal{P}_2(Q, \mu')\|_1 \leq (2 - \alpha + \lambda)\|\mu - \mu'\|_1$$

The function $(Q, \mu) \mapsto \mathcal{T}_2(Q, \mu)$ is Lipschitz with respect to both Q and μ:

$$\|\mathcal{T}_2(Q, \mu) - \mathcal{T}_2(Q', \mu)\|_\infty \leq (\gamma + 1)\|Q - Q'\|_\infty$$
$$\|\mathcal{T}_2(Q, \mu) - \mathcal{T}_2(Q, \mu')\|_\infty \leq (L_f + \gamma L_p\|Q\|_\infty)\|\mu - \mu'\|_1$$

Global Asymptotically Stable Equilibrium (GASE)

Under our assumptions, for any given μ, the ODE

$$\dot{Q}_t = \frac{1}{\epsilon}\mathcal{T}_2(Q_t, \mu)$$

has a unique GASE denoted by Q^*_μ. Moreover, Q^*_μ is uniformly Lipchitz in μ:

$$\|Q^*_\mu - Q^*_{\mu'}\|_\infty \leq \frac{L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty}{1-\gamma} \|\mu - \mu'\|_1$$

Additionally, one can show that the ODE

$$\dot{\mu}_t = \mathcal{P}_2(Q^*_{\mu_t}, \mu_t)$$

has a unique GASE denoted by $\mu^{*\phi}$.

Global Asymptotically Stable Equilibrium (GASE)

Under our assumptions, for any given μ, the ODE

$$\dot{Q}_t = \frac{1}{\epsilon}\mathcal{T}_2(Q_t, \mu)$$

has a unique GASE denoted by Q^*_μ. Moreover, Q^*_μ is uniformly Lipchitz in μ:

$$\|Q^*_\mu - Q^*_{\mu'}\|_\infty \leq \frac{L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty}{1-\gamma} \|\mu - \mu'\|_1$$

Additionally, one can show that the ODE

$$\dot{\mu}_t = \mathcal{P}_2(Q^*_{\mu_t}, \mu_t)$$

has a unique GASE denoted by $\mu^{*\phi}$.

Convergence

Assume that the learning rates ρ_n^Q and ρ_n^μ are sequences of positive real numbers satisfying

$$\sum_n \rho_n^Q = \sum_n \rho_n^\mu = \infty, \qquad \sum_n |\rho_n^Q|^2 + |\rho_n^\mu|^2 < \infty, \qquad \rho_n^\mu / \rho_n^Q \xrightarrow[n\to+\infty]{} 0$$

Typically $\rho_n^\mu = \frac{1}{(1+n)^{\omega_\mu}}$ and $\rho_n^Q = \frac{1}{(1+n)^{\omega_Q}}$ with $\frac{1}{2} < \omega_Q < \omega_\mu < 1$

Theorem

Following Borkar, the solution (μ_n, Q_n) to the idealized learning iterations

$$\begin{cases} \mu_{n+1} = \mu_n + \rho_n^\mu \mathcal{P}_2(Q_n, \mu_n) \\ Q_{n+1} = Q_n + \rho_n^Q \mathcal{T}_2(Q_n, \mu_n) \end{cases}$$

converges as $n \to \infty$ to $(\mu^{*\phi}, Q^*_{\mu^{*\phi}})$.

Convergence

Assume that the learning rates ρ_n^Q and ρ_n^μ are sequences of positive real numbers satisfying

$$\sum_n \rho_n^Q = \sum_n \rho_n^\mu = \infty, \qquad \sum_n |\rho_n^Q|^2 + |\rho_n^\mu|^2 < \infty, \qquad \rho_n^\mu / \rho_n^Q \xrightarrow[n\to+\infty]{} 0$$

Typically $\rho_n^\mu = \frac{1}{(1+n)^{\omega_\mu}}$ and $\rho_n^Q = \frac{1}{(1+n)^{\omega_Q}}$ with $\frac{1}{2} < \omega_Q < \omega_\mu < 1$

Theorem

Following Borkar, the solution (μ_n, Q_n) to the idealized learning iterations

$$\begin{cases} \mu_{n+1} = \mu_n + \rho_n^\mu \mathcal{P}_2(Q_n, \mu_n) \\ Q_{n+1} = Q_n + \rho_n^Q \mathcal{T}_2(Q_n, \mu_n) \end{cases}$$

converges as $n \to \infty$ to $(\mu^{*\phi}, Q^*_{\mu^{*\phi}})$.

Construction of the MFG solution

From $(\mu^{*\phi}, Q^*_{\mu^{*\phi}})$ obtained as the limit of the algorithm, we define:

1. Set $\pi^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$
2. Compute μ^* by using $\mu^* = \mu^* \mathrm{P}^{\pi^*, \mu^*}$
3. For fix μ^*, compute Q^* by using the Bellman equation.

Remark: If $\text{argmin-e}\, Q^*_{\mu^{*\phi}} = \text{argmin-e}\, Q^*$, then (μ^*, π^*) is a solution to the MFG problem. In general, we have the following closedness result:

Theorem

Let $\delta(\phi)$ be the **action gap** *defined as*
$\delta(\phi) = \min_{x\in\mathcal{X}}(\min_{a\notin \arg\min_a Q^*_{\mu^{*\phi}}} Q^*_{\mu^{*\phi}}(x,a) - \min_a Q^*_{\mu^{*\phi}}(x,a)) > 0$, *and*
$\delta(\phi) = \infty$ *if* $Q^*_{\mu^{*\phi}}(x)$ *is constant with respect to* a *for each* x. *Then:*

$$\begin{cases} \|\mu^{*\phi} - \mu^*\|_1 \leq \frac{2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{\alpha - \lambda} \\ \|Q^*_{\mu^{*\phi}} - Q^*\|_\infty \leq \frac{(L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty) 2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{(1-\gamma)(\alpha-\lambda)}. \end{cases}$$

Construction of the MFG solution

From $(\mu^{*\phi}, Q^*_{\mu^{*\phi}})$ obtained as the limit of the algorithm, we define:

1. Set $\pi^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$
2. Compute μ^* by using $\mu^* = \mu^* \mathrm{P}^{\pi^*, \mu^*}$
3. For fix μ^*, compute Q^* by using the Bellman equation.

Remark: If $\text{argmin-e}\, Q^*_{\mu^{*\phi}} = \text{argmin-e}\, Q^*$, then (μ^*, π^*) is a solution to the MFG problem.In general, we have the following closedness result:

Theorem

Let $\delta(\phi)$ be the **action gap** *defined as*
$\delta(\phi) = \min_{x\in\mathcal{X}}(\min_{a\notin \arg\min_a Q^*_{\mu^{*\phi}}} Q^*_{\mu^{*\phi}}(x,a) - \min_a Q^*_{\mu^{*\phi}}(x,a)) > 0$, *and*
$\delta(\phi) = \infty$ *if* $Q^*_{\mu^{*\phi}}(x)$ *is constant with respect to* a *for each* x. *Then:*

$$\begin{cases} \|\mu^{*\phi} - \mu^*\|_1 \leq \frac{2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{\alpha-\lambda} \\ \|Q^*_{\mu^{*\phi}} - Q^*\|_\infty \leq \frac{(L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty) 2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{(1-\gamma)(\alpha-\lambda)}. \end{cases}$$

Main Result: Identification of the Limit

Theorem

Assuming additionally that $\delta = \liminf_{\phi \to \infty} \delta(\phi) > 0$, *the distribution and policy* (μ^*, π^*) *form a solution to the MFG problem.*

Proof. From the idealized algorithm and by definition of π^*, we have:

$$Q^*_{\mu^{*\phi}}(x,a) = f(x,a,\mu^{*\phi}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}) \min_{a'} Q^*_{\mu^{*\phi}}(x',a')$$

$$Q^*_{\mu^{*\phi}}(x,a) > \min_{a'} Q^*_{\mu^{*\phi}}(x,a') \quad \text{for all} \quad a \notin \text{support}(\pi^*(x)).$$

The difference in the inequality is greater than the gap $\delta(\phi)$. By choosing ϕ large enough, we can replace $Q^*_{\mu^{*\phi}}$ by Q^* and $\mu^{*\phi}$ by μ^* and obtain: for all $a \notin \text{support}(\pi^*(x))$,

$$Q^*(x,a) = f(x,a,\mu^*) + \gamma \sum_{x'} p(x'|x,a,\mu^*) \min_{a'} Q^*(x',a') > \min_{a'} Q^*(x,a').$$

Combined with the assumed uniqueness of Q^*, we have $\text{argmin-e}\, Q^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$, so that (μ^*, π^*) is indeed an MFG solution.

Main Result: Identification of the Limit

Theorem

Assuming additionally that $\delta = \liminf_{\phi\to\infty} \delta(\phi) > 0$, *the distribution and policy* (μ^*, π^*) *form a solution to the MFG problem.*

Proof. From the idealized algorithm and by definition of π^*, we have:

$$Q^*_{\mu^{*\phi}}(x,a) = f(x,a,\mu^{*\phi}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}) \min_{a'} Q^*_{\mu^{*\phi}}(x',a')$$

$$Q^*_{\mu^{*\phi}}(x,a) > \min_{a'} Q^*_{\mu^{*\phi}}(x,a') \quad \text{for all} \quad a \notin \text{support}(\pi^*(x)).$$

The difference in the inequality is greater than the gap $\delta(\phi)$. By choosing ϕ large enough, we can replace $Q^*_{\mu^{*\phi}}$ by Q^* and $\mu^{*\phi}$ by μ^* and obtain: for all $a \notin \text{support}(\pi^*(x))$,

$$Q^*(x,a) = f(x,a,\mu^*) + \gamma \sum_{x'} p(x'|x,a,\mu^*) \min_{a'} Q^*(x',a') > \min_{a'} Q^*(x,a').$$

Combined with the assumed uniqueness of Q^*, we have $\text{argmin-e}\, Q^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$, so that (μ^*, π^*) is indeed an MFG solution.

Main Result: Identification of the Limit

Theorem

Assuming additionally that $\delta = \liminf_{\phi \to \infty} \delta(\phi) > 0$, *the distribution and policy* (μ^*, π^*) *form a solution to the MFG problem.*

Proof. From the idealized algorithm and by definition of π^*, we have:

$$Q^*_{\mu^{*\phi}}(x,a) = f(x,a,\mu^{*\phi}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}) \min_{a'} Q^*_{\mu^{*\phi}}(x',a')$$

$$Q^*_{\mu^{*\phi}}(x,a) > \min_{a'} Q^*_{\mu^{*\phi}}(x,a') \quad \text{for all} \quad a \notin \text{support}(\pi^*(x)).$$

The difference in the inequality is greater than the gap $\delta(\phi)$. By choosing ϕ large enough, we can replace $Q^*_{\mu^{*\phi}}$ by Q^* and $\mu^{*\phi}$ by μ^* and obtain: for all $a \notin \text{support}(\pi^*(x))$,

$$Q^*(x,a) = f(x,a,\mu^*) + \gamma \sum_{x'} p(x'|x,a,\mu^*) \min_{a'} Q^*(x',a') > \min_{a'} Q^*(x,a').$$

Combined with the assumed uniqueness of Q^*, we have $\text{argmin-e}\, Q^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$, so that (μ^*, π^*) is indeed an MFG solution.

Main Result: Identification of the Limit

Theorem

Assuming additionally that $\delta = \liminf_{\phi\to\infty} \delta(\phi) > 0$, *the distribution and policy* (μ^*, π^*) *form a solution to the MFG problem.*

Proof. From the idealized algorithm and by definition of π^*, we have:

$$Q^*_{\mu^{*\phi}}(x,a) = f(x,a,\mu^{*\phi}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}) \min_{a'} Q^*_{\mu^{*\phi}}(x',a')$$

$$Q^*_{\mu^{*\phi}}(x,a) > \min_{a'} Q^*_{\mu^{*\phi}}(x,a') \quad \text{for all} \quad a \notin \text{support}(\pi^*(x)).$$

The difference in the inequality is greater than the gap $\delta(\phi)$. By choosing ϕ large enough, we can replace $Q^*_{\mu^{*\phi}}$ by Q^* and $\mu^{*\phi}$ by μ^* and obtain: for all $a \notin \text{support}(\pi^*(x))$,

$$Q^*(x,a) = f(x,a,\mu^*) + \gamma \sum_{x'} p(x'|x,a,\mu^*) \min_{a'} Q^*(x',a') > \min_{a'} Q^*(x,a').$$

Combined with the assumed uniqueness of Q^*, we have $\text{argmin-e}\, Q^* = \text{argmin-e}\, Q^*_{\mu^{*\phi}}$, so that (μ^*, π^*) is indeed an MFG solution.

Mean Field Control Problem

For a given policy π and a state-action pair (x, a) we define the **modified policy**:

$$\tilde{\pi}^{(x,a)}(x') := \begin{cases} \delta_a & \text{if} \quad x' = x \\ \pi(x) & \text{for} \quad x' \neq x \end{cases}$$

Then, the **modified Q-function** is given by

$$Q^{\pi}(x, a) := f(x, a, \mu^{\tilde{\pi}^{(x,a)}}) + \mathbb{E}\left[\sum_{n \geq 1} \gamma^n f(X_n^{\pi}, A_n, \mu^{\pi})\right],$$

where $\mu^{\tilde{\pi}^{(x,a)}}$ is the limiting distribution of $X_n^{\tilde{\pi}^{(x,a)}}$, and where the expectation is along the Markovian dynamics:

$$X_0^{\pi} = x, A_0 = a, \quad A_n \sim \pi(X_n^{\pi}), \quad X_{n+1}^{\pi} \sim p(\cdot | X_n^{\pi}, A_n, \mu^{\pi})$$

The optimal Q-function is defined as:

$$Q^*(x, a) := \inf_{\pi} Q^{\pi}(x, a).$$

Mean Field Control Problem

For a given policy π and a state-action pair (x, a) we define the **modified policy**:

$$\tilde{\pi}^{(x,a)}(x') := \begin{cases} \delta_a & \text{if} \quad x' = x \\ \pi(x) & \text{for} \quad x' \neq x \end{cases}$$

Then, the **modified Q-function** is given by

$$Q^{\pi}(x, a) := f(x, a, \mu^{\tilde{\pi}^{(x,a)}}) + \mathbb{E}\left[\sum_{n \geq 1} \gamma^n f(X_n^{\pi}, A_n, \mu^{\pi})\right],$$

where $\mu^{\tilde{\pi}^{(x,a)}}$ is the limiting distribution of $X_n^{\tilde{\pi}^{(x,a)}}$, and where the expectation is along the Markovian dynamics:

$$X_0^{\pi} = x, A_0 = a, \quad A_n \sim \pi(X_n^{\pi}), \quad X_{n+1}^{\pi} \sim p(\cdot | X_n^{\pi}, A_n, \mu^{\pi})$$

The optimal Q-function is defined as:

$$Q^*(x, a) := \inf_{\pi} Q^{\pi}(x, a).$$

Mean Field Control Problem

For a given policy π and a state-action pair (x, a) we define the **modified policy**:

$$\tilde{\pi}^{(x,a)}(x') := \begin{cases} \delta_a & \text{if} \quad x' = x \\ \pi(x) & \text{for} \quad x' \neq x \end{cases}$$

Then, the **modified Q-function** is given by

$$Q^{\pi}(x, a) := f(x, a, \mu^{\tilde{\pi}^{(x,a)}}) + \mathbb{E}\left[\sum_{n \geq 1} \gamma^n f(X_n^{\pi}, A_n, \mu^{\pi})\right],$$

where $\mu^{\tilde{\pi}^{(x,a)}}$ is the limiting distribution of $X_n^{\tilde{\pi}^{(x,a)}}$, and where the expectation is along the Markovian dynamics:

$$X_0^{\pi} = x, A_0 = a, \quad A_n \sim \pi(X_n^{\pi}), \quad X_{n+1}^{\pi} \sim p(\cdot | X_n^{\pi}, A_n, \mu^{\pi})$$

The optimal Q-function is defined as:

$$Q^*(x, a) := \inf_{\pi} Q^{\pi}(x, a).$$

McKean–Vlasov Bellman Equation

We assume that there exists a unique policy π^* such that: $Q^*(x,a) := Q^{\pi^*}(x,a)$ for every (x,a). We also assume that for every $x \in \mathcal{X}$, $\{a \in \mathcal{A} : Q^*(x,a) = \min_{a'} Q^*(x,a')\}$ is a singleton, whose element is a pure control denoted by $\alpha^*(x) = \arg\min_{a'} Q^*(x,a') \in \mathcal{A}$.

In other words, the policy π^* is a pure control α^*

Then, Q^* satisfies the McKean–Vlasov Bellman equation:

$$Q^*(x,a) = f(x,a,\tilde{\mu}^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\tilde{\mu}^*) \min_{a'} Q^*(x',a'),$$

where $\tilde{\mu}^* := \mu^{\tilde{\pi}^*}$ is the invariant distribution associated to the policy $\tilde{\pi}^*$ which implicitly depends on (x,a).

And of course, in the RL algorithm, we choose $\rho_n^Q \ll \rho_n^\mu$

McKean–Vlasov Bellman Equation

We assume that there exists a unique policy π^* such that: $Q^*(x,a) := Q^{\pi^*}(x,a)$ for every (x,a). We also assume that for every $x \in \mathcal{X}$, $\{a \in \mathcal{A} : Q^*(x,a) = \min_{a'} Q^*(x,a')\}$ is a singleton, whose element is a pure control denoted by $\alpha^*(x) = \arg\min_{a'} Q^*(x,a') \in \mathcal{A}$.

In other words, the policy π^* is a pure control α^*

Then, Q^* satisfies the McKean–Vlasov Bellman equation:

$$Q^*(x,a) = f(x,a,\tilde{\mu}^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\tilde{\mu}^*) \min_{a'} Q^*(x',a'),$$

where $\tilde{\mu}^* := \mu^{\tilde{\pi}^*}$ is the invariant distribution associated to the policy $\tilde{\pi}^*$ which implicitly depends on (x,a).

And of course, in the RL algorithm, we choose $\rho_n^Q \ll \rho_n^\mu$

McKean–Vlasov Bellman Equation

We assume that there exists a unique policy π^* such that: $Q^*(x,a) := Q^{\pi^*}(x,a)$ for every (x,a). We also assume that for every $x \in \mathcal{X}$, $\{a \in \mathcal{A} : Q^*(x,a) = \min_{a'} Q^*(x,a')\}$ is a singleton, whose element is a pure control denoted by $\alpha^*(x) = \arg\min_{a'} Q^*(x,a') \in \mathcal{A}$.
In other words, the policy π^* is a pure control α^*
Then, Q^* satisfies the McKean–Vlasov Bellman equation:

$$Q^*(x,a) = f(x,a,\tilde{\mu}^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a,\tilde{\mu}^*) \min_{a'} Q^*(x',a'),$$

where $\tilde{\mu}^* := \mu^{\tilde{\pi}^*}$ is the invariant distribution associated to the policy $\tilde{\pi}^*$ which implicitly depends on (x,a).

And of course, in the RL algorithm, we choose $\rho_n^Q \ll \rho_n^\mu$

Two-timescale Mean Field Control Q-learning

Require: N_{steps}: number of steps; ϕ: parameter for soft-min policy; $(\rho^Q_{n,x,a})_{n,x,a}$: learning rates for the value function; $(\rho^\mu_{n,x,a})_{n,x,a}$: learning rates for the population distribution: MFC: $\rho^Q_{n,X_n,A_n} < \rho^\mu_{n,X_n,A_n}$

Initialization: $Q_0(x,a) = 0$ and $\mu_0^{(x,a)} = [\frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|}]$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

Observe $X_0^{(x,a)} \sim \mu_0^{(x,a)}$ for all (x,a)

Update mean fields: $\mu_0^{(x,a)} = \boldsymbol{\delta}(X_0^{(x,a)})$

for $n = 0, 1, 2, \ldots, N_{steps} - 1$ **do**

  Choose action: Choose $A_n^{(x,a)} = a$ if $X_n^{(x,a)} = x$, otherwise $A_n^{(x,a)} \sim \text{soft-min}_\phi\, Q_n(X_n^{(x,a)}, \cdot)$

  Observe the new states $X_{n+1}^{(x,a)} \sim p(\cdot | X_n^{(x,a)}, A_n^{(x,a)}, \mu_n^{(x,a)})$ **provided by the environment**

  Update population distributions: $\mu_{n+1}^{(x,a)} = \mu_n^{(x,a)} + \rho^\mu_{n,X_n^{(x,a)},A_n^{(x,a)}}(\boldsymbol{\delta}(X_{n+1}^{(x,a)}) - \mu_n^{(x,a)})$

  Observe the costs $f_{n+1}^{(x,a)} = f\left(X_n^{(x,a)}, A_n^{(x,a)}, \mu_n^{(x,a)}\right)$

  Update value function: If $X_n^{(x,a)} = x$,

$$Q_{n+1}(x,a) = Q_n(x,a) + \rho^Q_{n,X_n^{(x,a)},A_n^{(x,a)}} \left[f_{n+1}^{(x,a)} + \gamma \min_{a' \in \mathcal{A}} Q_n(X_{n+1}^{(x,a)}, a') - Q_n(x,a) \right]$$

  Otherwise, $Q_{n+1} = Q_n$

end for

Return $(\mu_{N_{steps}}^{(x,a)}, Q_{N_{steps}})$

Idealized Two-timescale Iteration Approach for MFC

The finite difference equation system for the MFC problem becomes:

$$\begin{cases} \mu_{n+1}^{(x,a)} = \mu_n^{(x,a)} + \rho_n^{\mu} \tilde{\mathcal{P}}_2^{(x,a)}(Q_n, \mu_n^{(x,a)}) \\ Q_{n+1}(x,a) = Q_n(x,a) + \rho_n^{Q} \mathcal{T}_2(Q_n, \mu_n^{(x,a)})(x,a) \end{cases}$$

where $\tilde{\mathcal{P}}_2^{(x,a)} : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \Delta^{|\mathcal{X}|}$ and $\mathcal{T}_2 : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$ are defined by:

$$\begin{cases} \tilde{\mathcal{P}}_2^{(x,a)}(Q,\mu)(y) = \sum_{x' \neq x} \mu(x') p(y|x', \text{soft-min}_\phi Q(x'), \mu) + \mu(x) p(y|x,a,\mu) - \mu(y) \\ \mathcal{T}_2(Q,\mu)(x,a) = f(x,a,\mu) + \gamma \sum_{x'} p(x'|x,a,\mu) \min_{a'} Q(x',a') - Q(x,a) \end{cases}$$

and where the first function can also be written as

$$\tilde{\mathcal{P}}_2^{(x,a)}(Q,\mu) = \mu \tilde{\mathrm{P}}_{(x,a)}^{\text{soft-min}_\phi Q, \mu} - \mu$$

which defines $\tilde{\mathrm{P}}_{(x,a)}^{\text{soft-min}_\phi Q, \mu}$

System of ODEs

Roughly speaking, in the regime $\rho_n^Q \ll \rho_n^\mu$, the solution $(\mu_n^{(x,a)}, Q_n)$ to the above iterations has a behavior that is described by the following system of ODEs, in the regime where ϵ is small and t goes to infinity:

$$\begin{cases} \dot{\mu}_t^{(x,a)} = \frac{1}{\epsilon}\tilde{\mathcal{P}}_2^{(x,a)}(Q_t, \mu_t^{(x,a)}) \\ \dot{Q}_t(x,a) = \mathcal{T}_2(Q_t, \mu_t^{(x,a)})(x,a) \end{cases}$$

where ρ_n^Q/ρ_n^μ is thought of being of order $\epsilon \ll 1$.

Similarly to the MFG case, we make Lipschitz assumptions on $\mathcal{P}_2^{(x,a)}$ so that the first equation as a unique GASE $\mu_Q^{*\phi}$ (omitting the dependency on (x,a)), and we assume that the second ODE along $\mu_Q^{*\phi}$ has a unique GASE denoted by $Q^{*\phi}$.

Let $V^{*\phi}(x) = \min_{a'} Q^{*\phi}(x,a')$ for all $x \in \mathcal{X}$ and define $V_n(x) = \min_{a'} Q_n(x,a')$ for all $x \in \mathcal{X}$, then we have the following convergence result:

Theorem

(μ_n, V_n) converges to $(\mu_{Q^{*\phi}}^{*\phi}, V^{*\phi})$ as $n \to \infty$, where we recall that μ_n and $\mu_{Q^{*\phi}}^{*\phi}$ depend on (x,a).

System of ODEs

Roughly speaking, in the regime $\rho_n^Q \ll \rho_n^\mu$, the solution $(\mu_n^{(x,a)}, Q_n)$ to the above iterations has a behavior that is described by the following system of ODEs, in the regime where ϵ is small and t goes to infinity:

$$\begin{cases} \dot{\mu}_t^{(x,a)} = \frac{1}{\epsilon}\tilde{\mathcal{P}}_2^{(x,a)}(Q_t, \mu_t^{(x,a)}) \\ \dot{Q}_t(x,a) = \mathcal{T}_2(Q_t, \mu_t^{(x,a)})(x,a) \end{cases}$$

where ρ_n^Q/ρ_n^μ is thought of being of order $\epsilon \ll 1$.

Similarly to the MFG case, we make Lipschitz assumptions on $\mathcal{P}_2^{(x,a)}$ so that the first equation as a unique GASE $\mu_Q^{*\phi}$ (omitting the dependency on (x,a)), and we assume that the second ODE along $\mu_Q^{*\phi}$ has a unique GASE denoted by $Q^{*\phi}$.

Let $V^{*\phi}(x) = \min_{a'} Q^{*\phi}(x,a')$ for all $x \in \mathcal{X}$ and define $V_n(x) = \min_{a'} Q_n(x,a')$ for all $x \in \mathcal{X}$, then we have the following convergence result:

Theorem

(μ_n, V_n) converges to $(\mu_{Q^{*\phi}}^{*\phi}, V^{*\phi})$ as $n \to \infty$, where we recall that μ_n and $\mu_{Q^{*\phi}}^{*\phi}$ depend on (x,a).

System of ODEs

Roughly speaking, in the regime $\rho_n^Q \ll \rho_n^\mu$, the solution $(\mu_n^{(x,a)}, Q_n)$ to the above iterations has a behavior that is described by the following system of ODEs, in the regime where ϵ is small and t goes to infinity:

$$\begin{cases} \dot{\mu}_t^{(x,a)} = \frac{1}{\epsilon}\tilde{\mathcal{P}}_2^{(x,a)}(Q_t, \mu_t^{(x,a)}) \\ \dot{Q}_t(x,a) = \mathcal{T}_2(Q_t, \mu_t^{(x,a)})(x,a) \end{cases}$$

where ρ_n^Q/ρ_n^μ is thought of being of order $\epsilon \ll 1$.
Similarly to the MFG case, we make Lipschitz assumptions on $\mathcal{P}_2^{(x,a)}$ so that the first equation as a unique GASE $\mu_Q^{*\phi}$ (omitting the dependency on (x,a)), and we assume that the second ODE along $\mu_Q^{*\phi}$ has a unique GASE denoted by $Q^{*\phi}$.
Let $V^{*\phi}(x) = \min_{a'} Q^{*\phi}(x,a')$ for all $x \in \mathcal{X}$ and define $V_n(x) = \min_{a'} Q_n(x,a')$ for all $x \in \mathcal{X}$, then we have the following convergence result:

Theorem

(μ_n, V_n) converges to $(\mu_{Q^{*\phi}}^{*\phi}, V^{*\phi})$ as $n \to \infty$, where we recall that μ_n and $\mu_{Q^{*\phi}}^{*\phi}$ depend on (x,a).

Construction of an MFC solution

1. Set $\alpha^*(x) = \arg\min_{a'} Q^{*\phi}(x, a')$
2. For each (x, a), define $\tilde{\alpha}^*$ and compute $\mu^{\tilde{\alpha}^*}$ solution to

$$\mu(y) = \sum_{x' \neq x} \mu(x') p(y|x', \alpha^*(x'), \mu) + \mu(x) p(y|x, a, \mu), \quad \forall y.$$

3. Compute V^{α^*} by using its definition (value of following α^* along $\mu^{\tilde{\alpha}^*}$)

The following result shows that $(\mu^{*\phi}_{Q^{*\phi}}, V^{*\phi})$ are close to $(\mu^{\tilde{\alpha}^*}, V^{\alpha^*})$

Theorem

Let $\delta(\phi)$ be the action gap defined as
$\delta(\phi) = \min_{x \in \mathcal{X}} (\min_{a \notin \arg\min_a Q^{*\phi}} Q^{*\phi}(x, a) - \min_a Q^{*\phi}(x, a)) > 0$, *and*
$\delta(\phi) = \infty$ *if $Q^{*\phi}(x)$ is constant with respect to a for each x. Then,*

$$\begin{cases} \|\mu^{*\phi}_{Q^{*\phi}} - \mu^*\|_1 \leq \frac{2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{\alpha - \lambda} \\ \|V^{*\phi} - V^{\alpha^*}\|_\infty \leq \frac{(L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty) 2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{(1-\gamma)(\alpha - \lambda)} \end{cases}$$

Construction of an MFC solution

1. Set $\alpha^*(x) = \arg\min_{a'} Q^{*\phi}(x, a')$
2. For each (x, a), define $\tilde{\alpha}^*$ and compute $\mu^{\tilde{\alpha}^*}$ solution to

$$\mu(y) = \sum_{x' \neq x} \mu(x') p(y|x', \alpha^*(x'), \mu) + \mu(x) p(y|x, a, \mu), \quad \forall y.$$

3. Compute V^{α^*} by using its definition (value of following α^* along $\mu^{\tilde{\alpha}^*}$)

The following result shows that $(\mu^{*\phi}_{Q^{*\phi}}, V^{*\phi})$ are close to $(\mu^{\tilde{\alpha}^*}, V^{\alpha^*})$

Theorem

Let $\delta(\phi)$ be the action gap defined as
$\delta(\phi) = \min_{x \in \mathcal{X}} (\min_{a \notin \arg\min_a Q^{*\phi}} Q^{*\phi}(x, a) - \min_a Q^{*\phi}(x, a)) > 0$, *and*
$\delta(\phi) = \infty$ *if* $Q^{*\phi}(x)$ *is constant with respect to* a *for each* x. *Then,*

$$\begin{cases} \|\mu^{*\phi}_{Q^{*\phi}} - \mu^*\|_1 \leq \frac{2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{\alpha - \lambda} \\ \|V^{*\phi} - V^{\alpha^*}\|_\infty \leq \frac{(L_f + \frac{\gamma}{1-\gamma} L_p \|f\|_\infty) 2|\mathcal{A}|^{\frac{3}{2}} \exp(-\phi\delta(\phi))}{(1-\gamma)(\alpha - \lambda)} \end{cases}$$

Characterization of the MFC Solution

$$V^*(x) := \inf_{\pi} V^{\pi}(x) = \min_{a} Q^*(x,a).$$

By taking $\delta_{\alpha^*(x)} = \pi^*(x)$ in the Bellman equation for Q^* one obtains the Bellman equation for $V^* = V^{\pi^*}$:

$$V^*(x) = f(x, \pi^*(x), \mu^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \pi^*(x), \mu^*) V^*(x'),$$

The solution α^* of the MFC problem can be characterized as follows:

1. Define Q^* by

$$Q^*(x,a) = Q^{\alpha^*}(x,a) = f(x, a, \mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x, a, \mu^{\tilde{\alpha}^*}) V^*(x').$$

2. Check that $\alpha^*(x) = \arg\min_a Q^*(x,a)$, so that Q^* satisfies the Bellman equation.

Characterization of the MFC Solution

$$V^*(x) := \inf_{\pi} V^{\pi}(x) = \min_{a} Q^*(x,a).$$

By taking $\delta_{\alpha^*(x)} = \pi^*(x)$ in the Bellman equation for Q^* one obtains the Bellman equation for $V^* = V^{\pi^*}$:

$$V^*(x) = f(x, \pi^*(x), \mu^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \pi^*(x), \mu^*) V^*(x'),$$

The solution α^* of the MFC problem can be characterized as follows:

1. Define Q^* by

$$Q^*(x,a) = Q^{\alpha^*}(x,a) = f(x,a,\mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x,a,\mu^{\tilde{\alpha}^*}) V^*(x').$$

2. Check that $\alpha^*(x) = \arg\min_a Q^*(x,a)$, so that Q^* satisfies the Bellman equation.

Final Result for MFC

Theorem

Assuming additionally that $\delta = \liminf_{\phi\to\infty} \delta(\phi) > 0$, *then the pure policy* α^* *is a solution to the MFC problem.*

Define Q^* by

$$Q^*(x,a) = Q^{\alpha^*}(x,a) = f(x,a,\mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x,a,\mu^{\tilde{\alpha}^*}) V^{\alpha^*}(x').$$

From our algorithm we have

$$f(x,a,\mu^{*\phi}_{Q^{*\phi}}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}_{Q^{*\phi}}) V^{*\phi}(x') > V^{*\phi}(x) \quad \text{for} \quad a \neq \alpha^*(x).$$

Using the positive gap condition $\delta(\phi) > 0$, by choosing ϕ large enough, we can replace $V^{*\phi}$ by V^{α^*} and $\mu^{*\phi}_{Q^{*\phi}}$ by $\mu^{\tilde{\alpha}^*}$ and obtain:

$$f(x,a,\mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x,a,\mu^{\tilde{\alpha}^*}) V^{\alpha^*}(x') > V^{\alpha^*}(x) \quad \text{for} \quad a \neq \alpha^*(x),$$

i.e. $Q^*(x,a) > V^*(x)$ for $a \neq \alpha^*(x)$ and therefore $\alpha^*(x) = \arg\min_a Q^*(x,a)$

Final Result for MFC

Theorem

Assuming additionally that $\delta = \liminf_{\phi\to\infty} \delta(\phi) > 0$, *then the pure policy* α^* *is a solution to the MFC problem.*

Define Q^* by

$$Q^*(x,a) = Q^{\alpha^*}(x,a) = f(x,a,\mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x,a,\mu^{\tilde{\alpha}^*}) V^{\alpha^*}(x').$$

From our algorithm we have

$$f(x,a,\mu^{*\phi}_{Q^{*\phi}}) + \gamma \sum_{x'} p(x'|x,a,\mu^{*\phi}_{Q^{*\phi}}) V^{*\phi}(x') > V^{*\phi}(x) \quad \text{for} \quad a \neq \alpha^*(x).$$

Using the positive gap condition $\delta(\phi) > 0$, by choosing ϕ large enough, we can replace $V^{*\phi}$ by V^{α^*} and $\mu^{*\phi}_{Q^{*\phi}}$ by $\mu^{\tilde{\alpha}^*}$ and obtain:

$$f(x,a,\mu^{\tilde{\alpha}^*}) + \gamma \sum_{x'} p(x'|x,a,\mu^{\tilde{\alpha}^*}) V^{\alpha^*}(x') > V^{\alpha^*}(x) \quad \text{for} \quad a \neq \alpha^*(x),$$

i.e. $Q^*(x,a) > V^*(x)$ for $a \neq \alpha^*(x)$ and therefore $\alpha^*(x) = \arg\min_a Q^*(x,a)$

From the idealized MFG algorithm to the synchronous stochastic approximation

The idealized algorithm relies on the operators $\mathcal{P}_2$ and $\mathcal{T}_2$ which involve expectations. Instead we sample $X'_{x,a,\mu} \sim p(\cdot|x,a,\mu)$, and we introduce

$$\begin{cases} \check{\mathcal{T}}_{\mu,Q}(x,a) = f(x,a,\mu) + \gamma \min_{a'} Q(X'_{x,a,\mu}, a') - Q(x,a) \\ (\check{\mathcal{P}}_{x,a}(\mu)(x'))_{x' \in \mathcal{X}} = \left(1_{\{X'_{x,a,\mu}=x'\}} - \mu(x')\right)_{x' \in \mathcal{X}} \end{cases}$$

centered at $\mathcal{P}_2$ and $\mathcal{T}_2$.

For MFG, starting from some initial Q_0, μ_0: for $n = 0, 1, ...,$

$$\begin{cases} \mu_{n+1}(x) & = \mu_n(x) + \rho_n^{\mu} \check{\mathcal{P}}_{X_n, \text{soft-min}\, Q(X_n)}(\mu_n)(x) \\ & = \mu_n(x) + \rho_n^{\mu}(\mathcal{P}_2(Q_n,\mu_n)(x) + \mathbf{P}_n(x)), \quad \forall x \in \mathcal{X} \\ Q_{n+1}(x,a) & = Q_n(x,a) + \rho_n^{Q} \check{\mathcal{T}}_{\mu_n,Q_n}(x,a) \\ & = Q_n(x,a) + \rho_n^{Q}(\mathcal{T}_2(Q_n,\mu_n)(x,a) + \mathbf{T}_n(x,a)), \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A} \\ X_n & \sim \mu_n \end{cases}$$

where $\mathbf{T}_n$ and $\mathbf{P}_n$ are **martingale difference sequences**.

Then apply classical **stochastic approximation**.

From the idealized MFG algorithm to the synchronous stochastic approximation

The idealized algorithm relies on the operators $\mathcal{P}_2$ and $\mathcal{T}_2$ which involve expectations. Instead we sample $X'_{x,a,\mu} \sim p(\cdot|x,a,\mu)$, and we introduce

$$\begin{cases} \check{\mathcal{T}}_{\mu,Q}(x,a) = f(x,a,\mu) + \gamma \min_{a'} Q(X'_{x,a,\mu}, a') - Q(x,a) \\ (\check{\mathcal{P}}_{x,a}(\mu)(x'))_{x' \in \mathcal{X}} = \left(1_{\{X'_{x,a,\mu}=x'\}} - \mu(x')\right)_{x' \in \mathcal{X}} \end{cases}$$

centered at $\mathcal{P}_2$ and $\mathcal{T}_2$.

For MFG, starting from some initial Q_0, μ_0: for $n = 0, 1, ...,$

$$\begin{cases} \mu_{n+1}(x) & = \mu_n(x) + \rho_n^{\mu} \check{\mathcal{P}}_{X_n, \text{soft-min}\, Q(X_n)}(\mu_n)(x) \\ & = \mu_n(x) + \rho_n^{\mu}(\mathcal{P}_2(Q_n,\mu_n)(x) + \mathbf{P}_n(x)), \quad \forall x \in \mathcal{X} \\ Q_{n+1}(x,a) & = Q_n(x,a) + \rho_n^{Q} \check{\mathcal{T}}_{\mu_n,Q_n}(x,a) \\ & = Q_n(x,a) + \rho_n^{Q}(\mathcal{T}_2(Q_n,\mu_n)(x,a) + \mathbf{T}_n(x,a)), \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A} \\ X_n & \sim \mu_n \end{cases}$$

where $\mathbf{T}_n$ and $\mathbf{P}_n$ are **martingale difference sequences**.

Then apply classical **stochastic approximation**.

From the idealized MFG algorithm to the synchronous stochastic approximation

The idealized algorithm relies on the operators $\mathcal{P}_2$ and $\mathcal{T}_2$ which involve expectations. Instead we sample $X'_{x,a,\mu} \sim p(\cdot|x,a,\mu)$, and we introduce

$$\begin{cases} \check{\mathcal{T}}_{\mu,Q}(x,a) = f(x,a,\mu) + \gamma \min_{a'} Q(X'_{x,a,\mu}, a') - Q(x,a) \\ (\check{\mathcal{P}}_{x,a}(\mu)(x'))_{x' \in \mathcal{X}} = \left(1_{\{X'_{x,a,\mu}=x'\}} - \mu(x')\right)_{x' \in \mathcal{X}} \end{cases}$$

centered at $\mathcal{P}_2$ and $\mathcal{T}_2$.

For MFG, starting from some initial Q_0, μ_0: for $n = 0, 1, ...,$

$$\begin{cases} \mu_{n+1}(x) & = \mu_n(x) + \rho_n^{\mu} \check{\mathcal{P}}_{X_n, \text{soft-min}\, Q(X_n)}(\mu_n)(x) \\ & = \mu_n(x) + \rho_n^{\mu}(\mathcal{P}_2(Q_n,\mu_n)(x) + \mathbf{P}_n(x)), \quad \forall x \in \mathcal{X} \\ Q_{n+1}(x,a) & = Q_n(x,a) + \rho_n^{Q} \check{\mathcal{T}}_{\mu_n,Q_n}(x,a) \\ & = Q_n(x,a) + \rho_n^{Q}(\mathcal{T}_2(Q_n,\mu_n)(x,a) + \mathbf{T}_n(x,a)), \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A} \\ X_n & \sim \mu_n \end{cases}$$

where $\mathbf{T}_n$ and $\mathbf{P}_n$ are **martingale difference sequences**.

Then apply classical **stochastic approximation**.

From the idealized MFC algorithm to the synchronous stochastic approximation

For MFC, we replace the deterministic updates by the following stochastic ones, starting from some initial Q_0, $\mu_0^{(x,a)}$: for $n = 0, 1, ...,$

$$\begin{cases} \mu_{n+1}^{(x,a)}(y) & = \mu_n^{(x,a)}(y) + \rho_n^{\mu} \check{\mathcal{P}}_{X_n^{(x,a)}, \tilde{\pi}(X_n^{(x,a)})}(\mu_n^{(x,a)})(y) \\ & = \mu_n^{(x,a)}(y) + \rho_n^{\mu}(\tilde{\mathcal{P}}_2^{(x,a)}(Q_n, \mu_n^{(x,a)})(y) + \mathbf{P}_n^{(x,a)}(y)), \quad \forall y \\ Q_{n+1}(x,a) & = Q_n(x,a) + \rho_n^{Q} \check{\mathcal{T}}_{\mu_n^{(x,a)}, Q_n}(x,a) \\ & = Q_n(x,a) + \rho_n^{Q}(\mathcal{T}_2(Q_n, \mu_n^{(x,a)})(x,a) + \mathbf{T}_n(x,a)), \quad \forall (x,a) \\ X_n^{(x,a)} & \sim \mu_n^{(x,a)} \end{cases}$$

From synchronous algorithm to the asynchronous full RL algorithm

In the **MFG asynchronous setting**, we consider the system

$$\mu_{n+1} = \mu_n + \rho^{\mu}_{n,X_n,A_n}(\mathcal{P}_2(Q_n, \mu_n) + \mathbf{P}_n)$$
$$Q_{n+1}(X_n, A_n) = Q_n(X_n, A_n) + \rho^{Q}_{n,X_n,A_n}(\mathcal{T}_2(Q_n, \mu_n)(X_n, A_n) + \mathbf{T}_n(X_n, A_n))$$

for $x \in \mathcal{X}$, $a \in \mathcal{A}, n \geq 0$, where (X_n, A_n) is the state-action pair at time n.

For MFC, we define the following system

$$\mu^{(x,a)}_{n+1} = \mu^{(x,a)}_n + \rho^{\mu}_{n,X^{(x,a)}_n,A^{(x,a)}_n}(\tilde{\mathcal{P}}_2(Q_n, \mu^{(x,a)}_n) + \mathbf{P}^{(x,a)}_n)$$
$$Q_{n+1}(x, a) = Q_n(x, a) + \rho^{Q}_{n,X^{(x,a)}_n,A^{(x,a)}_n}(\mathcal{T}_2(Q_n, \mu^{(x,a)}_n)(x, a) + \mathbf{T}_n(x, a)1_{\{X^{(x,a)}_n = x\}}$$

for $x \in \mathcal{X}$, $a \in \mathcal{A}, n \geq 0$, where $(X^{(x,a)}_n, A^{(x,a)}_n)$ is the state-action pair at time n.

From synchronous algorithm to the asynchronous full RL algorithm

In the **MFG asynchronous setting**, we consider the system

$$\mu_{n+1} = \mu_n + \rho^{\mu}_{n,X_n,A_n}(\mathcal{P}_2(Q_n, \mu_n) + \mathbf{P}_n)$$
$$Q_{n+1}(X_n, A_n) = Q_n(X_n, A_n) + \rho^{Q}_{n,X_n,A_n}(\mathcal{T}_2(Q_n, \mu_n)(X_n, A_n) + \mathbf{T}_n(X_n, A_n))$$

for $x \in \mathcal{X}$, $a \in \mathcal{A}, n \geq 0$, where (X_n, A_n) is the state-action pair at time n.

For MFC, we define the following system

$$\mu^{(x,a)}_{n+1} = \mu^{(x,a)}_n + \rho^{\mu}_{n,X^{(x,a)}_n,A^{(x,a)}_n}(\tilde{\mathcal{P}}_2(Q_n, \mu^{(x,a)}_n) + \mathbf{P}^{(x,a)}_n)$$
$$Q_{n+1}(x, a) = Q_n(x, a) + \rho^{Q}_{n,X^{(x,a)}_n,A^{(x,a)}_n}(\mathcal{T}_2(Q_n, \mu^{(x,a)}_n)(x, a) + \mathbf{T}_n(x, a)1_{\{X^{(x,a)}_n = x\}}$$

for $x \in \mathcal{X}$, $a \in \mathcal{A}, n \geq 0$, where $(X^{(x,a)}_n, A^{(x,a)}_n)$ is the state-action pair at time n.

Learning rates and trajectory properties

We use the following **learning rates**:

$$\rho^Q_{n,x,a} = \frac{1}{(1+\nu(x,a,n))^{\omega^Q}}, \qquad \rho^\mu_n = \frac{1}{(1+n)^{\omega^\mu}},$$

where $\nu(x,a,n)$ is the number of times that the algorithm visited state x and performed action a until time t_n.
The exponent ω^Q can take values in $(\frac{1}{2}, 1)$.
The value of ω^μ is chosen depending on the value of ω^Q and the **MFG (competitive)** or **MFC (cooperative)** nature of the problem we want to solve.
These learning rates satisfy the **Ideal tapering stepsize (ITS)** properties related to the (almost sure) good behavior of the number visits along trajectories.
Additionally: there exists a deterministic $\Delta > 0$ such that for all (x,a), $\liminf_{n\to\infty} \frac{\nu(x,a,n)}{n} \geq \Delta \quad a.s.$

Learning rates and trajectory properties

We use the following **learning rates**:

$$\rho^Q_{n,x,a} = \frac{1}{(1+\nu(x,a,n))^{\omega^Q}}, \qquad \rho^\mu_n = \frac{1}{(1+n)^{\omega^\mu}},$$

where $\nu(x,a,n)$ is the number of times that the algorithm visited state x and performed action a until time t_n.
The exponent ω^Q can take values in $(\frac{1}{2}, 1)$.
The value of ω^μ is chosen depending on the value of ω^Q and the **MFG (competitive)** or **MFC (cooperative)** nature of the problem we want to solve.
These learning rates satisfy the **Ideal tapering stepsize (ITS)** properties related to the (almost sure) good behavior of the number visits along trajectories.
Additionally: there exists a deterministic $\Delta > 0$ such that for all (x,a), $\liminf_{n\to\infty} \frac{\nu(x,a,n)}{n} \geq \Delta \quad a.s.$

Extension to Mean Field Control Games (MFCG)

Competitive game between a large number of large **collaborative groups** of agents in the infinite limit of number and size of groups.

$$f(x, a, \mu, \mu'), \quad p(\cdot \mid x, a, \mu, \mu')$$

where μ **is the global population distribution** and μ' **is the local population distribution**.

For instance, from a classical MFG setup $f(x, a, \mu)$ and $p(\cdot \mid x, a, \mu)$, introduce some **degree λ of collaboration**:

$$f(x, a, (1-\lambda)\mu + \lambda\mu'), \quad p(\cdot \mid x, a, (1-\lambda)\mu + \lambda\mu'))$$

The algorithm becomes three-timescale with

$$\rho^{\mu} \ll \rho^{Q} \ll \rho^{\mu'}$$

A. Angiuli, J.-P. Fouque, M. Lauriere, and M. Zhang: Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games, **Applied Mathematics & Optimization, 2026** (arxiv.org/abs/2405.17017)

Extension to Mean Field Control Games (MFCG)

Competitive game between a large number of large **collaborative groups** of agents in the infinite limit of number and size of groups.

$$f(x, a, \mu, \mu'), \quad p(\cdot \mid x, a, \mu, \mu')$$

where μ **is the global population distribution** and μ' **is the local population distribution**.

For instance, from a classical MFG setup $f(x, a, \mu)$ and $p(\cdot \mid x, a, \mu)$, introduce some **degree λ of collaboration**:

$$f(x, a, (1-\lambda)\mu + \lambda\mu'), \quad p(\cdot \mid x, a, (1-\lambda)\mu + \lambda\mu'))$$

The algorithm becomes three-timescale with

$$\rho^{\mu} \ll \rho^{Q} \ll \rho^{\mu'}$$

A. Angiuli, J.-P. Fouque, M. Lauriere, and M. Zhang: Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games, **Applied Mathematics & Optimization, 2026** (arxiv.org/abs/2405.17017)

Extension to Mean Field Control Games (MFCG)

Competitive game between a large number of large **collaborative groups** of agents in the infinite limit of number and size of groups.

$$f(x, a, \mu, \mu'), \quad p(\cdot \mid x, a, \mu, \mu')$$

where μ **is the global population distribution** and μ' **is the local population distribution**.

For instance, from a classical MFG setup $f(x, a, \mu)$ and $p(\cdot \mid x, a, \mu)$, introduce some **degree λ of collaboration**:

$$f(x, a, (1-\lambda)\mu + \lambda\mu'), \quad p(\cdot \mid x, a, (1-\lambda)\mu + \lambda\mu'))$$

The algorithm becomes three-timescale with

$$\rho^{\mu} \ll \rho^{Q} \ll \rho^{\mu'}$$

A. Angiuli, J.-P. Fouque, M. Lauriere, and M. Zhang: Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games, **Applied Mathematics & Optimization, 2026** (arxiv.org/abs/2405.17017)

Extension to Mean Field Control Games (MFCG)

Competitive game between a large number of large **collaborative groups** of agents in the infinite limit of number and size of groups.

$$f(x, a, \mu, \mu'), \quad p(\cdot \mid x, a, \mu, \mu')$$

where μ **is the global population distribution** and μ' **is the local population distribution**.

For instance, from a classical MFG setup $f(x, a, \mu)$ and $p(\cdot \mid x, a, \mu)$, introduce some **degree λ of collaboration**:

$$f(x, a, (1-\lambda)\mu + \lambda\mu'), \quad p(\cdot \mid x, a, (1-\lambda)\mu + \lambda\mu'))$$

The algorithm becomes three-timescale with

$$\rho^{\mu} \ll \rho^{Q} \ll \rho^{\mu'}$$

A. Angiuli, J.-P. Fouque, M. Lauriere, and M. Zhang: Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games, **Applied Mathematics & Optimization, 2026** (arxiv.org/abs/2405.17017)

MFCG 3-timescale Q-learning Algorithm

Require: N_{steps}: number of steps; ϕ: parameter for soft-min policy; $(\rho^Q_{n,x,a})$: learning rates for the value function; (ρ^μ_n): learning rate for the global distribution; $(\rho^{\tilde{\mu}}_{n,x,a})$: learning rates for the local distribution.

Initialization: $Q_0(x,a) = 0$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\mu_0 = \mu_0^{(x,a)} = [\frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|}]$

Observe $X_0 \sim \mu_0$, $X_0^{(x,a)} \sim \mu_0^{(x,a)}$

Update distributions: $\mu_0 = \boldsymbol{\delta}(X_0)$, $\mu_0^{(x,a)} = \boldsymbol{\delta}(X_0^{(x,a)})$

for $n = 0, 1, 2, \ldots, N_{steps} - 1$ **do**

  Choose action $A_n \sim \text{soft-min}_\phi\, Q_n(X_n, \cdot)$ and observe state $X_{n+1} \sim p(\cdot | X_n, A_n, \mu_n, \mu_n^{(X_n,A_n)})$ and cost $f_{n+1} = f(X_n, A_n, \mu_n, \mu_n^{(X_n,A_n)})$ provided by the environment

  Choose action Choose $A_n^{(x,a)} = a$ if $X_n^{(x,a)} = x$, otherwise $A_n^{(x,a)} \sim \text{soft-min}_\phi\, Q_n(X_n^{(x,a)}, \cdot)$ and observe state $X_{n+1}^{(x,a)} \sim p(\cdot | X_n^{(x,a)}, A_n^{(x,a)}, \mu_n, \mu_n^{(x,a)})$

  Update local distributions: $\mu_{n+1}^{(x,a)} = \mu_n^{(x,a)} + \rho^{\tilde{\mu}}_{n, X_n^{(x,a)}, A_n^{(x,a)}} (\boldsymbol{\delta}(X_{n+1}^{(x,a)}) - \mu_n^{(x,a)})$

  Update global distribution: $\mu_{n+1} = \mu_n + \rho^\mu_n (\boldsymbol{\delta}(X_{n+1}) - \mu_n)$

  if $X_n^{(X_n,A_n)} = X_n$ **then**

    Update value function: $Q_{n+1}(x,a) = Q_n(x,a)$ for all $(x,a) \neq (X_n, A_n)$, and

$$Q_{n+1}(X_n, A_n) = Q_n(X_n, A_n) + \rho^Q_{n, X_n, A_n} [f_{n+1} + \gamma \min_{a' \in \mathcal{A}} Q_n(X_{n+1}, a') - Q_n(X_n, A_n)]$$

  end if

end for

Return $(\mu_{N_{steps}}, \mu^{(x,a)}_{N_{steps}}, Q_{N_{steps}})$

Infinite horizon paper:
https://arxiv.org/abs/2312.06659

Thank you all for your attention

Infinite horizon paper:
https://arxiv.org/abs/2312.06659

Thank you all for your attention