Reinforcement Learning for Mean Field Problems: a Unified Multi-Timescale Approach

Andrea Angiuli

Amazon Robotics
(work pre-dates current Amazon affiliation)

in collaboration with:
J.-P. Fouque, M. Laurière, R. Hu, A. Raydan, N. Detering, J. Lin, M. Zhang

Institute for Mathematical and Statistical Innovation (IMSI)
Chicago, 2026

Outline

Outline

The N-player game

N symmetric players with pairwise interactions. Player i's state $X_t^i \in \mathbb{R}^d$ depends on *every* other player:

$$dX_t^i = b\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dt + \sigma\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dW_t^i, \qquad X_t^{-i} = (X_t^j)_{j \neq i}.$$

$$J^i(\alpha) = \mathbb{E}\bigg[\int_0^\infty e^{-\beta t} f\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dt\bigg].$$

Curse of dimensionality. $O(N^2)$ pairwise interactions. Intractable for large N.

Symmetry hypothesis. b, σ, f are symmetric in the labels of X_t^{-i} (players are *indistinguishable*). Then the dependence on X_t^{-i} is approximated by a dependence on the **empirical distribution** $\bar{\boldsymbol{\mu}}_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{X_t^j}$:

$$dX_t^i = b\Big(t, X_t^i, \alpha_t^i, \bar{\boldsymbol{\mu}}_t^N\Big)\, dt + \sigma\Big(t, X_t^i, \alpha_t^i, \bar{\boldsymbol{\mu}}_t^N\Big)\, dW_t^i.$$

The N-player game

N symmetric players with pairwise interactions. Player i's state $X_t^i \in \mathbb{R}^d$ depends on *every* other player:

$$dX_t^i = b\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dt + \sigma\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dW_t^i, \qquad X_t^{-i} = (X_t^j)_{j \neq i}.$$

$$J^i(\alpha) = \mathbb{E}\bigg[\int_0^\infty e^{-\beta t} f\Big(t, X_t^i, \alpha_t^i, X_t^{-i}\Big)\, dt\bigg].$$

Curse of dimensionality. $O(N^2)$ pairwise interactions. Intractable for large N.

Symmetry hypothesis. b, σ, f are symmetric in the labels of X_t^{-i} (players are *indistinguishable*). Then the dependence on X_t^{-i} is approximated by a dependence on the **empirical distribution** $\bar{\boldsymbol{\mu}}_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{X_t^j}$:

$$dX_t^i = b\Big(t, X_t^i, \alpha_t^i, \bar{\boldsymbol{\mu}}_t^N\Big)\, dt + \sigma\Big(t, X_t^i, \alpha_t^i, \bar{\boldsymbol{\mu}}_t^N\Big)\, dW_t^i.$$

The mean field limit

Take $N \to \infty$. The system of N players collapses to the dynamics of a representative agent X_t interacting with the distribution of the population:

$$dX_t = b(t, X_t, \alpha_t, \mu_t)dt + \sigma(t, X_t, \alpha_t, \mu_t)dW_t, \qquad \mu_t = \mathcal{L}(X_t).$$

Two solution concepts in the limit, by nature of the interaction:

- **Competitive** $\longrightarrow$ *Mean Field Game* (MFG): extension of *Nash equilibrium*.
- **Cooperative** $\longrightarrow$ *Mean Field Control* (MFC): extension of *social optimum*.

The mean field limit

Take $N \to \infty$. The system of N players collapses to the dynamics of a representative agent X_t interacting with the distribution of the population:

$$dX_t = b(t, X_t, \alpha_t, \mu_t)dt + \sigma(t, X_t, \alpha_t, \mu_t)dW_t, \qquad \mu_t = \mathcal{L}(X_t).$$

Two solution concepts in the limit, by nature of the interaction:

- **Competitive** $\longrightarrow$ *Mean Field Game* (MFG): extension of *Nash equilibrium*.
- **Cooperative** $\longrightarrow$ *Mean Field Control* (MFC): extension of *social optimum*.

The mean field limit: MFG and MFC (asymptotic, infinite horizon)

Take $N \to \infty$ and look at a representative agent. State distribution μ.

Mean Field Game (Nash)

Find $(\hat{\alpha}, \hat{\mu})$ such that

1. $\hat{\alpha} = \arg\min_\alpha J^{AMFG}(\alpha; \hat{\mu})$, with $\hat{\mu}$ frozen;
2. $\hat{\mu} = \lim_{n\to\infty} \mathcal{L}\left(X_n^{\hat{\alpha},\hat{\mu}}\right)$.

Infinitesimal player joining a crowd already at equilibrium.

Mean Field Control (social optimum)

Find α^* minimizing

$$J^{AMFC}(\alpha) = \mathbb{E}\Big[\sum_{n\geq 0} \gamma^n f(X_n^\alpha, \alpha(X_n^\alpha), \mu^\alpha)\Big],$$

with $\mu^\alpha = \lim_n \mathcal{L}(X_n^\alpha)$ moving with α.

Central planner picking the best stationary distribution.

Why reinforcement learning?

Existing numerical methods (PDE, FBSDE, deep solvers) require knowing the model (b, σ, f).

Many real applications don't. Banking, energy markets, crowd dynamics, opinion propagation: the agent only sees transitions and rewards.

Goal of this talk. Algorithms that learn $(\hat{\alpha}, \hat{\mu})$ *or* (α^*, μ^*)

- from samples,
- model-free,
- from a **single representative agent** who does **not** observe the population's distribution.

μ is estimated from the representative agent's trajectory, exploiting symmetry / limiting problem.

Outline

The MDP and 1-step Temporal Difference (TD)

The Markov decision process. An agent observes X_n, takes A_n, gets cost f_{n+1}, transitions to X_{n+1}. Aggregate (discounted) cost $G_n = \sum_{k\geq 0} \gamma^k f_{n+k+1}$.

All updates have the form $V(X_n) \leftarrow V(X_n) + \rho\,[T_n - V(X_n)]$, that differ based on the *target* T_n:

Method	**Target T_n**	**Properties**
Monte Carlo	$G_n = \sum_{k\geq 0} \gamma^k f_{n+k+1}$	Unbiased; high variance; requires the complete return (episode termination or truncation).
1-step TD	$f_{n+1} + \gamma\, V(X_{n+1})$	Biased (bootstraps); low variance; *online*.
n-step TD	$\sum_{j=0}^{k-1} \gamma^j f_{n+j+1} + \gamma^k V(X_{n+k})$	Trades bias and variance.

Stochastic approximation as a dynamical system

Robbins–Monro (1951). Find the root of an unknown $h : \mathbb{R}^d \to \mathbb{R}^d$ from noisy samples

$$\theta_{k+1} = \theta_k + \rho_k \left[h(\theta_k) + M_{k+1} \right],$$

with $\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = 0$ – martingale-difference due to sampling noise. Under $\sum_k \rho_k = \infty,\ \sum_k \rho_k^2 < \infty$, the iterates **track the ODE** $\dot{\theta}_t = h(\theta_t)$.

Convergence question, restated. Does $\dot{\theta} = h(\theta)$ admit a **globally asymptotically stable equilibrium** (GASE)? If yes, $\theta_k \to \theta^\infty$ a.s.

The rest of the talk

Every iteration we will see — Q-table, distribution μ, actor, critic, score function — will be analyzed as an ODE tracked by a stochastic update.

Stochastic approximation as a dynamical system

Robbins–Monro (1951). Find the root of an unknown $h : \mathbb{R}^d \to \mathbb{R}^d$ from noisy samples

$$\theta_{k+1} = \theta_k + \rho_k \left[h(\theta_k) + M_{k+1} \right],$$

with $\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = 0$ – martingale-difference due to sampling noise. Under $\sum_k \rho_k = \infty,\ \sum_k \rho_k^2 < \infty$, the iterates **track the ODE** $\dot{\theta}_t = h(\theta_t)$.

Convergence question, restated. Does $\dot{\theta} = h(\theta)$ admit a **globally asymptotically stable equilibrium** (GASE)? If yes, $\theta_k \to \theta^\infty$ a.s.

The rest of the talk

Every iteration we will see — Q-table, distribution μ, actor, critic, score function — will be analyzed as an ODE tracked by a stochastic update.

Two-timescale stochastic approximation

Coupled iterations on $(\theta, w) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$:

$$\theta_{k+1} = \theta_k + \rho_k^\theta \left[H(\theta_k, w_k) + M_{k+1}^\theta \right],$$
$$w_{k+1} = w_k + \rho_k^w \left[G(\theta_k, w_k) + M_{k+1}^w \right],$$

with **separated rates**: $\rho_k^w / \rho_k^\theta \to 0$ as $k \to \infty$.

Effective dynamics (Borkar 1997). The slow loop *sees* the fast one as already at its equilibrium $\theta(w)$, the fast loop *sees* w as frozen. The system tracks

$$\dot{\theta}_t \;=\; \tfrac{1}{\epsilon}\, H(\theta_t, w_t)\,, \qquad \dot{w}_t \;=\; G(\theta(w_t), w_t)\,, \qquad \epsilon \to 0.$$

Borkar–Konda (1997): actor-critic.

- Critic (value evaluation under fixed policy) = fast; its ODE has GASE V_π.
- Actor (policy improvement using the converged critic) = slow.

Two-timescale stochastic approximation

Coupled iterations on $(\theta, w) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$:

$$\theta_{k+1} = \theta_k + \rho_k^\theta \left[H(\theta_k, w_k) + M_{k+1}^\theta \right],$$
$$w_{k+1} = w_k + \rho_k^w \left[G(\theta_k, w_k) + M_{k+1}^w \right],$$

with **separated rates**: $\rho_k^w / \rho_k^\theta \to 0$ as $k \to \infty$.

Effective dynamics (Borkar 1997). The slow loop *sees* the fast one as already at its equilibrium $\theta(w)$, the fast loop *sees* w as frozen. The system tracks

$$\dot{\theta}_t \;=\; \tfrac{1}{\epsilon}\, H(\theta_t, w_t)\,, \qquad \dot{w}_t \;=\; G(\theta(w_t), w_t)\,, \qquad \epsilon \to 0.$$

Borkar–Konda (1997): actor-critic.

- Critic (value evaluation under fixed policy) = fast; its ODE has GASE V_π.
- Actor (policy improvement using the converged critic) = slow.

RL for Mean Field Problems

Look at MFG/MFC through the same lens.

- **Agent side**: Q-function, updated by Bellman residuals *given the current* μ.
- **Population side**: distribution μ, updated by the law of states under the current policy.

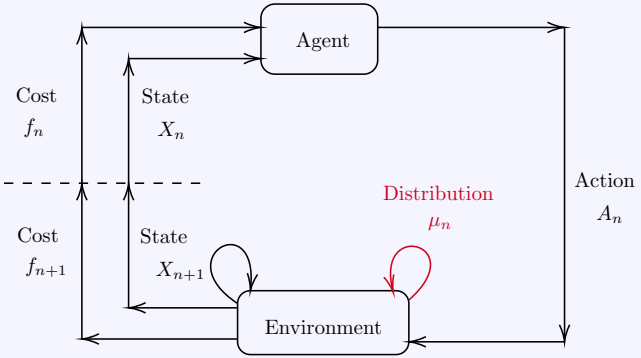
Two natural orderings of timescales.

Q rate	μ rate	Limit problem
fast	slow	**MFG** (frozen $\hat{\mu}$)
slow	fast	**MFC** (μ co-moves with α)

Outline

Setting: an MDP with a mean-field parameter

Finite state $\mathcal{X}$, finite action $\mathcal{A}$, dynamics $p(x' \mid x, a, \mu)$, one-step cost $f(x, a, \mu)$, discount $\gamma \in (0, 1)$.



Key restriction. The agent does not know p, f and does not observe μ. The environment estimates μ from the agent's own trajectory — exploiting symmetry: at equilibrium, $\mathcal{L}(X) = \mu$.

Q-function for asymptotic MFG: a frozen-μ MDP

Fix any $\mu \in \mathcal{P}(\mathcal{X})$ and consider the standard Q-function of an infinitesimal agent against this μ:

$$Q^*_\mu(x,a) \;=\; \min_\alpha \mathbb{E}\Big[\textstyle\sum_{n\geq 0}\gamma^n f(X_n,\alpha(X_n),\mu) \;\big|\; X_0{=}x, A_0{=}a\Big].$$

Since μ is frozen this is a **classical MDP**. By dynamic programming, Q^*_μ solves the Bellman equation

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

Equilibrium: $\hat\alpha(x) = \arg\min_a Q^*_{\hat\mu}(x,a)$ and $\hat\mu = \lim_n \mathcal{L}\left(X_n^{\hat\alpha,\hat\mu}\right)$ (fixed point).

Q-function for asymptotic MFG: a frozen-μ MDP

Fix any $\mu \in \mathcal{P}(\mathcal{X})$ and consider the standard Q-function of an infinitesimal agent against this μ:

$$Q^*_\mu(x,a) \;=\; \min_\alpha \mathbb{E}\Big[\textstyle\sum_{n\geq 0}\gamma^n f(X_n,\alpha(X_n),\mu) \;\big|\; X_0{=}x, A_0{=}a\Big].$$

Since μ is frozen this is a **classical MDP**. By dynamic programming, Q^*_μ solves the Bellman equation

$$\boxed{Q^*_\mu(x,a) \;=\; f(x,a,\mu) \;+\; \gamma\sum_{x'} p(x' \mid x,a,\mu)\,\min_{a'} Q^*_\mu(x',a').}$$

Equilibrium: $\hat\alpha(x) = \arg\min_a Q^*_{\hat\mu}(x,a)$ and $\hat\mu = \lim_n \mathcal{L}\left(X_n^{\hat\alpha,\hat\mu}\right)$ (fixed point).

Q-function for asymptotic MFG: a frozen-μ MDP

Fix any $\mu \in \mathcal{P}(\mathcal{X})$ and consider the standard Q-function of an infinitesimal agent against this μ:

$$Q^*_\mu(x,a) \;=\; \min_\alpha \mathbb{E}\Big[\textstyle\sum_{n\geq 0}\gamma^n f(X_n,\alpha(X_n),\mu) \;\big|\; X_0{=}x, A_0{=}a\Big].$$

Since μ is frozen this is a **classical MDP**. By dynamic programming, Q^*_μ solves the Bellman equation

$$\boxed{Q^*_\mu(x,a) \;=\; f(x,a,\mu) \;+\; \gamma\sum_{x'} p(x' \mid x,a,\mu)\, \min_{a'} Q^*_\mu(x',a').}$$

Equilibrium: $\hat\alpha(x) = \arg\min_a Q^*_{\hat\mu}(x,a)$ and $\hat\mu = \lim_n \mathcal{L}\left(X_n^{\hat\alpha,\hat\mu}\right)$ (fixed point).

Q-function for asymptotic MFC: a *modified* Bellman

In MFC, μ moves with α; the standard construction breaks.

For an admissible α and a given (x, a), define

$$\widetilde{\alpha}(x') \;=\; \begin{cases} a & x' = x, \\ \alpha(x') & x' \neq x. \end{cases}$$

Let $\widetilde{\mu} = \mu^{\widetilde{\alpha}}$ be the resulting limit distribution and define

$$Q^{\alpha}(x,a) \;=\; f(x,a,\widetilde{\mu}) + \mathbb{E}\Big[\textstyle\sum_{n\geq 1}\gamma^n f(X_n,\alpha(X_n),\mu^{\alpha})\,|X_0{=}x, A_0{=}a\Big].$$

Then $Q^* = \min_\alpha Q^\alpha$ satisfies the **modified Bellman**

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

with $\alpha^*(x) = \arg\min_a Q^*(x,a)$ and $\widetilde{\mu}^* = \mu^{\widetilde{\alpha}^*}$.

Q-function for asymptotic MFC: a *modified* Bellman

In MFC, μ moves with α; the standard construction breaks.

For an admissible α and a given (x, a), define

$$\widetilde{\alpha}(x') \;=\; \begin{cases} a & x' = x, \\ \alpha(x') & x' \neq x. \end{cases}$$

Let $\widetilde{\mu} = \mu^{\widetilde{\alpha}}$ be the resulting limit distribution and define

$$Q^{\alpha}(x,a) \;=\; f(x,a,\widetilde{\mu}) + \mathbb{E}\Big[\textstyle\sum_{n\geq 1}\gamma^n f(X_n,\alpha(X_n),\mu^{\alpha})\,|X_0{=}x, A_0{=}a\Big].$$

Then $Q^* = \min_\alpha Q^\alpha$ satisfies the **modified Bellman**

$$\boxed{Q^*(x,a) = f(x,a,\widetilde{\mu}^*) + \gamma \sum_{x'} p(x' \mid x, a, \widetilde{\mu}^*)\, \min_{a'} Q^*(x',a'),}$$

with $\alpha^*(x) = \arg\min_a Q^*(x,a)$ and $\widetilde{\mu}^* = \mu^{\widetilde{\alpha}^*}$.

Two coupled *stochastic* iterations

Idea. Update both Q and μ from samples, but at different rates. At each step, observe a transition $X' \sim p(\cdot \mid X, A, \mu)$ with $X \sim \mu$, $A = \arg\min_a Q(X, a)$, and form the sample estimators

$$\check{\mathcal{T}}_{Q,\mu,x,a} = f(x, a, \mu) + \gamma \min_{a'} Q(X', a') - Q(x, a),$$
$$\check{\mathcal{P}}_{Q,\mu,X,A}(x'') = \mathbb{1}_{\{X'=x''\}} - \mu(x'').$$

Then iterate, starting from (Q_0, μ_0):

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu} \check{\mathcal{P}}_{Q_k,\mu_k,X_k,A_k}, \\ Q_{k+1} = Q_k + \rho_k^{Q} \check{\mathcal{T}}_{Q_k,\mu_k,X_k,A_k}. \end{cases}$$

This is a Robbins–Monro scheme tracking

$$\dot{\mu}_t = \mathcal{P}(Q_t, \mu_t), \qquad \dot{Q}_t = \mathcal{T}(Q_t, \mu_t),$$

where $\mathcal{P}, \mathcal{T}$ are the (model-known) distribution-update and Bellman-residual operators.

Two coupled *stochastic* iterations

Idea. Update both Q and μ from samples, but at different rates. At each step, observe a transition $X' \sim p(\cdot \mid X, A, \mu)$ with $X \sim \mu$, $A = \arg\min_a Q(X, a)$, and form the sample estimators

$$\check{\mathcal{T}}_{Q,\mu,x,a} = f(x, a, \mu) + \gamma \min_{a'} Q(X', a') - Q(x, a),$$
$$\check{\mathcal{P}}_{Q,\mu,X,A}(x'') = \mathbb{1}_{\{X'=x''\}} - \mu(x'').$$

Then iterate, starting from (Q_0, μ_0):

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu} \check{\mathcal{P}}_{Q_k,\mu_k,X_k,A_k}, \\ Q_{k+1} = Q_k + \rho_k^{Q} \check{\mathcal{T}}_{Q_k,\mu_k,X_k,A_k}. \end{cases}$$

This is a Robbins–Monro scheme tracking

$$\dot{\mu}_t = \mathcal{P}(Q_t, \mu_t), \qquad \dot{Q}_t = \mathcal{T}(Q_t, \mu_t),$$

where $\mathcal{P}, \mathcal{T}$ are the (model-known) distribution-update and Bellman-residual operators.

Two coupled *stochastic* iterations

Idea. Update both Q and μ from samples, but at different rates. At each step, observe a transition $X' \sim p(\cdot \mid X, A, \mu)$ with $X \sim \mu$, $A = \arg\min_a Q(X, a)$, and form the sample estimators

$$\check{\mathcal{T}}_{Q,\mu,x,a} = f(x, a, \mu) + \gamma \min_{a'} Q(X', a') - Q(x, a),$$
$$\check{\mathcal{P}}_{Q,\mu,X,A}(x'') = \mathbb{1}_{\{X'=x''\}} - \mu(x'').$$

Then iterate, starting from (Q_0, μ_0):

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu} \check{\mathcal{P}}_{Q_k,\mu_k,X_k,A_k}, \\ Q_{k+1} = Q_k + \rho_k^{Q} \check{\mathcal{T}}_{Q_k,\mu_k,X_k,A_k}. \end{cases}$$

This is a Robbins–Monro scheme tracking

$$\dot{\mu}_t = \mathcal{P}(Q_t, \mu_t), \qquad \dot{Q}_t = \mathcal{T}(Q_t, \mu_t),$$

where $\mathcal{P}, \mathcal{T}$ are the (model-known) distribution-update and Bellman-residual operators.

One algorithm, two regimes

Tune the **ratio** ρ_k^μ / ρ_k^Q and apply Borkar's two-timescale theorem.

MFG regime: $\rho_k^\mu / \rho_k^Q \to 0$

Q is fast (μ frozen):
$\dot{Q}_t = \frac{1}{\epsilon}\mathcal{T}(Q_t, \mu) \to Q_\mu$.
Slow μ tracks $\dot{\mu}_t = \mathcal{P}(Q_{\mu_t}, \mu_t)$.
GASE at μ_∞ with $\mathcal{P}(Q_{\mu_\infty}, \mu_\infty) = 0 \Rightarrow$ Nash equilibrium.

MFC regime: $\rho_k^Q / \rho_k^\mu \to 0$

μ is fast (Q frozen):
$\dot{\mu}_t = \frac{1}{\epsilon}\mathcal{P}(Q, \mu_t) \to \mu_Q$.
Slow Q tracks $\dot{Q}_t = \mathcal{T}(Q_t, \widetilde{\mu}_{Q_t})$.
GASE at $Q_\infty = Q^* \Rightarrow$ MFC optimum via the modified Bellman.

Take-away

Same iterations, two ODE systems, two solutions. The choice of which timescale is fast *is* the choice of solution concept.

Algorithm: U2-MF-QL (tabular, asynchronous)

Algorithm 1 Unified Two-Timescale Mean Field Q-Learning (U2-MF-QL)

Input: horizon T, spaces $\mathcal{X}, \mathcal{A}$, init. μ_0, ϵ-greedy parameter, exponents (ω^Q, ω^μ).

Initialize $Q^0 \equiv 0$, $\mu_n^0 = [1/|\mathcal{X}|, \dots, 1/|\mathcal{X}|]$ for $n = 0, \dots, T$.
for episode $k = 1, 2, \dots$ **do**
  Sample $X_0^k \sim \mu_T^{k-1}$, set $Q^k \equiv Q^{k-1}$.
  for $n = 0, \dots, T-1$ **do**
    $\mu_n^k = \mu_n^{k-1} + \rho_k^\mu \left(\delta(X_n^k) - \mu_n^{k-1}\right)$ (slow μ)
    Choose A_n^k via ϵ-greedy on $Q^k(X_n^k, \cdot)$; observe f_{n+1}, X_{n+1}^k.
    $Q^k(X_n^k, A_n^k) \mathrel{+}= \rho^Q_{k,n,X_n^k,A_n^k} \left[f_{n+1} + \gamma \min_{a'} Q^k(X_{n+1}^k, a') - Q^k(X_n^k, A_n^k)\right]$ (fast Q)
  end for
end for
return (μ^k, Q^k), control $\widehat{\alpha}(x) = \arg\min_a Q^k(x, a)$.

Rates: $\rho^Q_{k,n,x,a} = (1 + \#|x, a, k, n|)^{-\omega^Q}$, $\rho_k^\mu = (1+k)^{-\omega^\mu}$, with $\omega^Q \in (\frac{1}{2}, 1)$.
$\omega^\mu > \omega^Q \Rightarrow$ MFG. $\omega^Q > \omega^\mu \Rightarrow$ MFC.

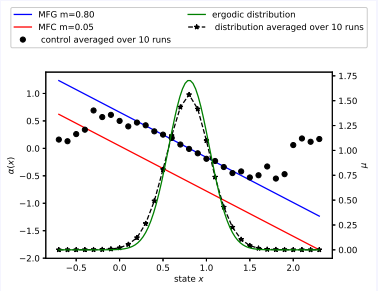
LQ benchmark

Continuous-state benchmark (discretized, $\Delta t = 10^{-2}$, $\sigma = 0.3$):

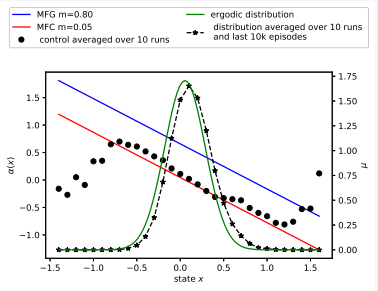
$$f(x,\alpha,\mu) = \tfrac{1}{2}\alpha^2 + c_1(x - c_2 m)^2 + c_3(x - c_4)^2 + c_5 m^2, \quad b(x,\alpha,\mu) = \alpha, \quad m = \int x\,\mu(dx).$$

Closed-form solutions: $\hat{m} = \frac{c_3 c_4}{c_1 + c_3 - c_1 c_2}$ (MFG), $m^* = \frac{c_3 c_4}{c_1 + c_3 + c_5 - c_1 c_2 (2 - c_2)}$ (MFC).

MFG: $(\omega^Q, \omega^\mu) = (0.55, 0.85)$



MFC: $(\omega^Q, \omega^\mu) = (0.65, 0.15)$



Same data $(c_1, c_2, c_3, c_4, c_5) = (0.25, 1.5, 0.5, 0.6, 5)$, same code, one parameter swap $\Rightarrow$ two distinct solutions.

Outline

From asymptotic to extended finite-horizon

Two generalizations of the previous setting:

1. **Finite horizon** $[0,T]$. Control becomes time-dependent: $\alpha_n(\cdot) := \alpha(n,\cdot)$.
2. **Extended interaction.** Population coupling through the *joint* state-action distribution $\nu_t = \mathcal{L}(X_t, \alpha_t(X_t))$, with first marginal μ_t and second marginal θ_t.

MFG: find $(\hat{\alpha}, \hat{\nu})$ such that

1. $\hat{\alpha} = \arg\min_\alpha \mathbb{E}\Big[\sum_{n=0}^{T-1} f(X_n^{\alpha,\hat{\nu}}, \alpha_n(X_n), \hat{\nu}_n) + g(X_T, \hat{\mu}_T)\Big]$;
2. $\hat{\nu}_n = \mathcal{L}(X_n^{\hat{\alpha}}, \hat{\alpha}_n(X_n^{\hat{\alpha}}))$, $n \in \{0,\dots,T-1\}$.

MFC: minimize the same objective with $\nu_n^\alpha = \mathcal{L}(X_n^\alpha, \alpha_n(X_n^\alpha))$ evolving with α.

U2-MF-QL-FH: time-indexed Q-table

New unknown. A 3-D table $\boldsymbol{Q} = (Q_n)_{n=0,\dots,N_T}$, $Q_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. At each episode k, time n, observed (X_n^k, A_n^k):

$$\nu_n^k = \nu_n^{k-1} + \rho_k^{\nu} \left(\delta(X_n^k, A_n^k) - \nu_n^{k-1} \right),$$

$$Q_n^k(x,a) \mathrel{+}= \mathbb{1}_{\{(x,a)=(X_n^k, A_n^k)\}} \, \rho_{x,a,k}^{Q_n} \left[\mathcal{B} - Q_n^{k-1}(x,a) \right],$$

$$\mathcal{B} = \begin{cases} f_{n+1} + \gamma \, \min_{a'} Q_{n+1}^{k-1}(X_{n+1}^k, a') & n < T, \\ f_{n+1} & n = T-1. \end{cases}$$

Asynchronous rates with per-time-point counters:
$\rho_{x,a,k}^{Q_n} = (1+N_T \, \#|x,a,t_n,k|)^{-\omega^Q}$, $\rho_k^{\nu} = (1+k)^{-\omega^{\nu}}$.

Same selection mechanism

$\omega^{\nu} > \omega^Q \Rightarrow$ MFG. $\omega^Q > \omega^{\nu} \Rightarrow$ MFC.

U2-MF-QL-FH: time-indexed Q-table

New unknown. A 3-D table $\boldsymbol{Q} = (Q_n)_{n=0,\dots,N_T}$, $Q_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. At each episode k, time n, observed (X_n^k, A_n^k):

$$\nu_n^k = \nu_n^{k-1} + \rho_k^{\nu} \left(\delta(X_n^k, A_n^k) - \nu_n^{k-1} \right),$$

$$Q_n^k(x,a) \mathrel{+}= \mathbb{1}_{\{(x,a)=(X_n^k, A_n^k)\}} \, \rho_{x,a,k}^{Q_n} \left[\mathcal{B} - Q_n^{k-1}(x,a) \right],$$

$$\mathcal{B} = \begin{cases} f_{n+1} + \gamma \, \min_{a'} Q_{n+1}^{k-1}(X_{n+1}^k, a') & n < T, \\ f_{n+1} & n = T-1. \end{cases}$$

Asynchronous rates with per-time-point counters:
$\rho_{x,a,k}^{Q_n} = (1+N_T \, \#|x,a,t_n,k|)^{-\omega^Q}$, $\rho_k^{\nu} = (1+k)^{-\omega^{\nu}}$.

Same selection mechanism

$\omega^{\nu} > \omega^Q \Rightarrow$ MFG. $\omega^Q > \omega^{\nu} \Rightarrow$ MFC.

The HARA capital-accumulation problem [Huang 2013]

Wealth of the representative agent (no borrowing):

$$X_{t+1}^{\alpha,\theta} \;=\; G\big(\textstyle\int a\, d\theta_t(a),\; W_t\big)\;\alpha_t, \qquad \alpha_t \in [0, X_t^{\alpha,\theta}].$$

G is a production function depending on the population's mean investment $z_t = \int a\, d\theta_t(a)$ and noise W_t. Consumption $c_t = X_t^{\alpha,\theta} - \alpha_t$.

Hyperbolic Absolute Risk Aversion (HARA) utility, finite horizon. Maximize

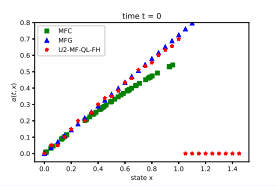
$$J(\alpha,\theta) \;=\; \mathbb{E}\Big[\textstyle\sum_{t=0}^{T} \rho^t\, \frac{1}{\gamma}\left(X_t^{\alpha,\theta} - \alpha_t\right)^{\gamma}\Big], \qquad \rho \in (0,1],\ \gamma \in (0,1).$$

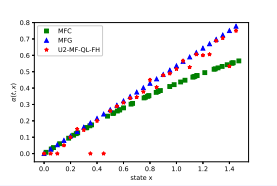
Why this is a good showcase:

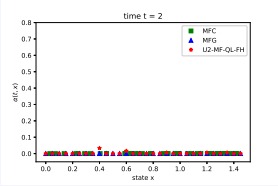
- Discrete time, *finite horizon*;
- *Multiplicative* noise $G(z, W)$;
- Interaction through the *control* distribution θ;
- Closed-form MFG solution; deep solver (Carmona–Laurière, 2019) as MFC benchmark.

HARA – learned controls at $t = 0, 1, 2$

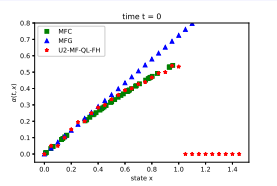
MFG: $(\omega^Q, \omega^\theta) = (0.55, 0.85)$

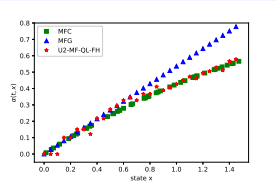


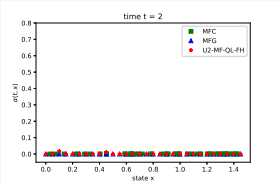




MFC: $(\omega^Q, \omega^\theta) = (0.7, 0.05)$



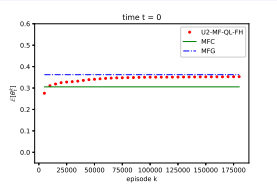


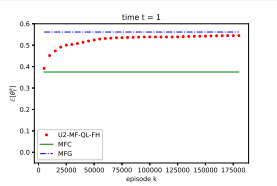


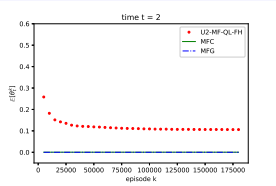
Red dots: learned controls (avg. over 10 runs). **Blue / green markers:** closed-form MFG / NN-benchmark MFC.

HARA – learned mean of the control distribution $\mathbb{E}[\alpha_t]$

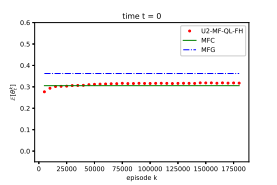
MFG: $(\omega^Q, \omega^\theta) = (0.55, 0.85)$

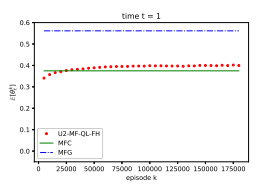


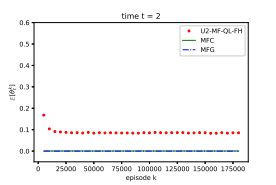




MFC: $(\omega^Q, \omega^\theta) = (0.7, 0.05)$







Convergence of $\mathbb{E}[\alpha_t^k]$ vs. episode k, for each $t \in \{0, 1, 2\}$. **The algorithm learns the right θ, not just the right α.**

Outline

The continuous-space challenge

What we have so far. Tabular Q-learning on $\mathcal{X}, \mathcal{A}$ finite. For continuous problems we have been discretizing $(\mathcal{X}, \mathcal{A})$ on a grid.

Why this breaks. In dimension d:

- $|\mathcal{X}| = O(N^d)$ grid points $\Rightarrow$ *curse of dimensionality*.
- Storing μ as a histogram on a fine grid is infeasible.
- Boundary truncation introduces artifacts.

Replace tables by neural networks.

Tabular	$\rightarrow$	Deep parametric
Q-table $Q(x,a)$	$\rightarrow$	critic V_θ + actor Π_ψ
histogram $\mu(x)$	$\rightarrow$	Stein score $\Sigma_\varphi(x) \approx \nabla \log p_\mu(x)$

Goal. *Same* unified-rate principle, now at the level of three (or four) neural networks.

Actor-critic, in one slide

Bellman-residual TD error. $\delta_n = r_{n+1} + \gamma V_\theta(X_{n+1}) - V_\theta(X_n)$.

Critic update (policy evaluation):

$$L_V^{(n)}(\theta) = \delta_n^2, \qquad \theta \leftarrow \theta + 2\rho_V \, \delta_n \, \nabla_\theta V_\theta(X_n).$$

Actor update (policy gradient theorem with TD-error baseline):

$$L_\Pi^{(n)}(\psi) = -\delta_n \, \log \Pi_\psi(A_n \,|\, X_n), \qquad \psi \leftarrow \psi + \rho_\Pi \, \delta_n \, \nabla_\psi \log \Pi_\psi(A_n \,|\, X_n).$$

Standard rate ordering: $\rho_\Pi < \rho_V$. Critic fast (policy evaluation must be accurate before policy improvement); actor slow.
This is exactly Borkar–Konda (1997).

Actor-critic, in one slide

Bellman-residual TD error. $\delta_n = r_{n+1} + \gamma V_\theta(X_{n+1}) - V_\theta(X_n)$.

Critic update (policy evaluation):

$$L_V^{(n)}(\theta) = \delta_n^2, \qquad \theta \leftarrow \theta + 2\rho_V \, \delta_n \, \nabla_\theta V_\theta(X_n).$$

Actor update (policy gradient theorem with TD-error baseline):

$$L_\Pi^{(n)}(\psi) = -\delta_n \, \log \Pi_\psi(A_n \,|\, X_n), \qquad \psi \leftarrow \psi + \rho_\Pi \, \delta_n \, \nabla_\psi \log \Pi_\psi(A_n \,|\, X_n).$$

Standard rate ordering: $\rho_\Pi < \rho_V$**.** Critic fast (policy evaluation must be accurate before policy improvement); actor slow.
This is exactly Borkar–Konda (1997).

Score-matching and Langevin sampling

Represent μ via its Stein score $s_\mu(x) := \nabla \log p_\mu(x)$, approximated by a neural network $\Sigma_\varphi : \mathbb{R}^d \to \mathbb{R}^d$.

Hyvärinen identity. Minimizing $\mathbb{E}_{x\sim\mu} \|\Sigma_\varphi(x) - s_\mu(x)\|^2$ is, by integration by parts, equivalent to minimizing

$$\boxed{\mathbb{E}_{x\sim\mu}\Big[\mathrm{tr}(\nabla_x \Sigma_\varphi(x)) + \tfrac{1}{2}\,\|\Sigma_\varphi(x)\|^2\Big].}$$

No knowledge of s_μ needed.

Online step at time n:

$$L_\Sigma^{(n)}(\varphi) = \mathrm{tr}(\nabla_x \Sigma_\varphi(X_n)) + \tfrac{1}{2}\,\|\Sigma_\varphi(X_n)\|^2, \quad \varphi \leftarrow \varphi - \rho_\Sigma\, \nabla_\varphi L_\Sigma^{(n)}(\varphi).$$

Sampling from μ via Langevin dynamics:

$$x_{m+1} = x_m + \tfrac{\epsilon}{2}\,\Sigma_\varphi(x_m) + \sqrt{\epsilon}\, z_m, \qquad z_m \sim \mathcal{N}(0, I).$$

As $\epsilon \to 0$, $m \to \infty$: $x_m \to_d \mu$.

Algorithm: IH-MF-AC (3-timescale actor-critic)

Algorithm 2 Infinite Horizon Mean Field Actor-Critic (IH-MF-AC)

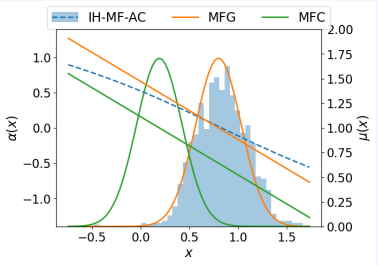
Input: step size Δt, learning rates $(\rho_\Pi, \rho_V, \rho_\Sigma)$, Langevin step ϵ.

Initialize networks: actor Π_{ψ_0}, critic V_{θ_0}, score Σ_{φ_0}.
Sample initial state $X_0 \sim \mu_0$.
for $n = 0, \dots, N-1$ **do**
    Update score φ: SGD step on $L_\Sigma^{(n)}$.
    Sample $S_n^{(1)}, \dots, S_n^{(K)}$ from $\Sigma_{\varphi_{n+1}}$ via Langevin $\Rightarrow$ $\overline{\mu}_{S_n}$.
    Sample action $A_n \sim \Pi_{\psi_n}(\cdot \mid X_n)$.
    Observe $r_{n+1} = -f(X_n, \overline{\mu}_{S_n}, A_n)\Delta t$, X_{n+1}.
    Compute TD error $\delta_n = r_{n+1} + e^{-\beta \Delta t} V_{\theta_n}(X_{n+1}) - V_{\theta_n}(X_n)$.
    Critic step: $\theta \leftarrow \theta + 2\rho_V\, \delta_n\, \nabla_\theta V_\theta(X_n)$.
    Actor step: $\psi \leftarrow \psi + \rho_\Pi\, \delta_n\, \nabla_\psi \log \Pi_\psi(A_n \mid X_n)$.
end for
return $(\Pi_{\psi_N}, \Sigma_{\varphi_N})$.

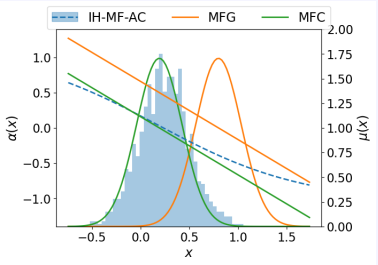
1D LQ benchmark in continuous space

Same LQ cost as before, now solved **without discretizing** $\mathcal{X}, \mathcal{A}$.

MFG: $\rho_\Sigma < \min(\rho_\Pi, \rho_V)$



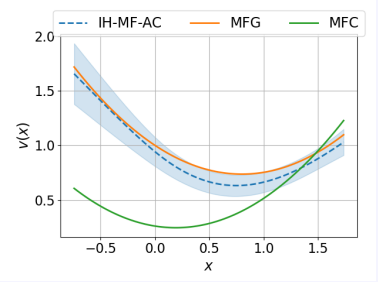
MFC: $\rho_\Sigma > \max(\rho_\Pi, \rho_V)$



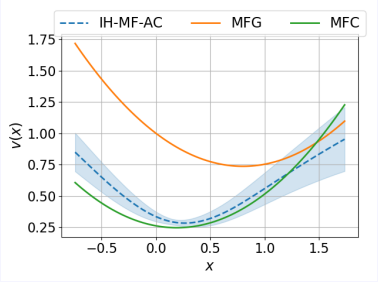
Histogram (blue): learned μ from Langevin samples of Σ_φ. Dashed line (blue): learned control. Orange / green: closed-form MFG / MFC.

1D LQ benchmark: value function

MFG: $\rho_\Sigma < \min(\rho_\Pi, \rho_V)$



MFC: $\rho_\Sigma > \max(\rho_\Pi, \rho_V)$



Dashed blue: learned value function $-V_{\theta_N}$ after $N = 10^6$ iterations, averaged over 5 runs (shaded ± 1 std).
Orange / green: closed-form MFG / MFC.

Three small NNs ($\sim$ 128 neurons each) recover both regimes from the *same* algorithm, by tuning the rate ordering.

Outline

From MFG/MFC to MFCG: cooperation *and* competition

Many real systems are not purely competitive nor purely cooperative.

Motivation. M groups of N agents each. Agent (m, n) is the n-th member of group m.

- **Within group** m: agents *cooperate* $\Rightarrow$ MFC-like.
- **Across groups**: groups *compete* $\Rightarrow$ MFG-like.

Mean field limit. Take $M, N \to \infty$. The result is a **Mean Field Control Game (MFCG)** described from the perspective of a *representative group*, with two distributions:

$$\mu_t \text{ (\textit{global}: full population)}, \qquad \widetilde{\mu}_t = \mathcal{L}(X_t^{\hat{\alpha},\mu}) \text{ (\textit{local}: representative group)}.$$

New solution concept. A pair $(\hat{\alpha}, \hat{\mu})$ where

1. the representative group solves a McKean–Vlasov **control** problem with the global $\hat{\mu}$ *frozen*;
2. fixed-point: $\mathcal{L}(X_t^{\hat{\alpha},\hat{\mu}}) = \hat{\mu}_t$.

Three-timescale Q-learning for MFCG

Three coupled iterations – one Q, two distributions:

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu}\, \mathcal{P}(Q_k, \mu_k), & \text{(global, slow)} \\ Q_{k+1} = Q_k + \rho_k^{Q}\, \mathcal{T}(Q_k, \mu_k, \widetilde{\mu}_k), & \text{(Q, intermediate)} \\ \widetilde{\mu}_{k+1} = \widetilde{\mu}_k + \rho_k^{\widetilde{\mu}}\, \mathcal{P}(Q_k, \widetilde{\mu}_k), & \text{(local, fast)} \end{cases}$$

with rate ordering $\rho_k^{\mu} < \rho_k^{Q} < \rho_k^{\widetilde{\mu}}$, i.e. $\omega^{\mu} > \omega^{Q} > \omega^{\widetilde{\mu}}$.

Why three timescales?

- Global μ is *frozen* during the agent's optimization (across-group MFG) $\Rightarrow$ must be the slowest.
- Q uses the modified Bellman from §3 (one local $\mu^*_{x,a}$ per state-action).
- $\widetilde{\mu}$ *moves with* α (within-group MFC) $\Rightarrow$ must be the fastest.

Three-timescale Q-learning for MFCG

Three coupled iterations – one Q, two distributions:

$$\begin{cases} \mu_{k+1} = \mu_k + \rho_k^{\mu}\, \mathcal{P}(Q_k, \mu_k), & \text{(global, slow)} \\ Q_{k+1} = Q_k + \rho_k^{Q}\, \mathcal{T}(Q_k, \mu_k, \widetilde{\mu}_k), & \text{(Q, intermediate)} \\ \widetilde{\mu}_{k+1} = \widetilde{\mu}_k + \rho_k^{\widetilde{\mu}}\, \mathcal{P}(Q_k, \widetilde{\mu}_k), & \text{(local, fast)} \end{cases}$$

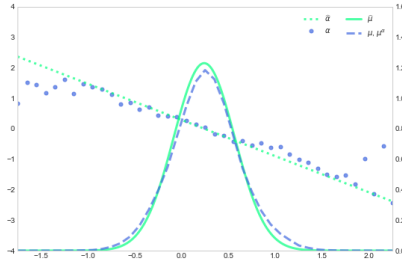
with rate ordering $\rho_k^{\mu} < \rho_k^{Q} < \rho_k^{\widetilde{\mu}}$, i.e. $\omega^{\mu} > \omega^{Q} > \omega^{\widetilde{\mu}}$.

Why three timescales?

- Global μ is *frozen* during the agent's optimization (across-group MFG) $\Rightarrow$ must be the slowest.
- Q uses the modified Bellman from §3 (one local $\mu^*_{x,a}$ per state-action).
- $\widetilde{\mu}$ *moves with* α (within-group MFC) $\Rightarrow$ must be the fastest.

LQ MFCG benchmark

Same LQ cost as in §3, augmented with a within-group penalty $\widetilde{c}_1(x - \widetilde{c}_2\bar{\widetilde{\mu}})^2 + \widetilde{c}_3(\bar{\widetilde{\mu}} - \widetilde{c})^2$.



Black: closed-form MFCG. Coloured: learned solution with $\rho^{\mu} < \rho^{Q} < \rho^{\widetilde{\mu}}$.

Other MFCG applications

The same three-timescale algorithm has been applied to:

- **Trader's optimal liquidation between competing groups of cooperative traders**
  Angiuli–Detering–Fouque–Laurière–Lin, arXiv:2205.02330.
- **Intra- and inter-bank borrowing and lending**
  extension of Carmona–Fouque–Sun (2015) systemic-risk model;
  Angiuli–Detering–Fouque–Laurière–Lin, arXiv:2207.03449.

Each application produces an LQ-type benchmark; in every case the three-timescale algorithm approximates the closed form.

MFCG goes deep: actor-critic with two score networks

Drop in the deep AC machinery from §5: replace

- global histogram $\mu \to$ global score net Σ_φ,
- local histogram $\widetilde{\mu} \to$ local score net $\widetilde{\Sigma}_\xi$.

Four learning rates, one ordering:

$$\rho_\Sigma \;<\; \min(\rho_\Pi, \rho_V) \;\le\; \max(\rho_\Pi, \rho_V) \;<\; \rho_{\widetilde{\Sigma}}.$$

Closes the loop.

Setting	# rates	Selects between
Tabular Q (U2-MF-QL)	2	MFG, MFC
Three-timescale Q (U3-MF-QL)	3	MFG, MFC, MFCG
Deep AC (IH-MF-AC)	3	MFG, MFC
Deep AC + 2 scores (IH-MFCG-AC)	4	MFG, MFC, MFCG

MFCG goes deep: actor-critic with two score networks

Drop in the deep AC machinery from §5: replace

- global histogram $\mu \to$ global score net Σ_φ,
- local histogram $\widetilde{\mu} \to$ local score net $\widetilde{\Sigma}_\xi$.

Four learning rates, one ordering:

$$\rho_\Sigma \;<\; \min(\rho_\Pi, \rho_V) \;\le\; \max(\rho_\Pi, \rho_V) \;<\; \rho_{\widetilde{\Sigma}}.$$

Closes the loop.

Setting	# rates	Selects between
Tabular Q (U2-MF-QL)	2	MFG, MFC
Three-timescale Q (U3-MF-QL)	3	MFG, MFC, MFCG
Deep AC (IH-MF-AC)	3	MFG, MFC
Deep AC + 2 scores (IH-MFCG-AC)	4	MFG, MFC, MFCG

Outline

Conclusion and open directions

Take-away.

- One principle – **separate the timescales of agent-side and population-side updates**.
- Two rates select MFG vs. MFC; three rates handle MFCG.
- Same recipe scales from tabular Q to deep AC + score-matching.
- Model-free: one representative agent, no oracle on μ.

Open directions.

- Convergence in continuous spaces (deep AC + score-matching).
- More applications: market making, energy markets, opinion dynamics.

Thank you. Questions?

References

1. A. Angiuli, J.-P. Fouque, M. Laurière. **Unified Reinforcement Q-Learning for Mean Field Game and Control Problems**. *Mathematics of Control, Signals, and Systems*, Vol. 34, p. 217–271, 2022.
2. A. Angiuli, J.-P. Fouque, M. Laurière. **Reinforcement Learning for Mean Field Games, with Applications to Economics**. In: *Machine Learning and Data Sciences for Financial Markets* (A. Capponi, C.-A. Lehalle, eds.), Cambridge University Press, 2023.
3. A. Angiuli, N. Detering, J.-P. Fouque, M. Laurière, J. Lin. **Reinforcement Learning Algorithm for Mixed Mean Field Control Games**. *Journal of Machine Learning*, Vol. 2, p. 108–137, 2023.
4. A. Angiuli, N. Detering, J.-P. Fouque, M. Laurière, J. Lin. **Reinforcement Learning for Intra- and Inter-Bank Borrowing and Lending Mean Field Control Game**. *3rd ACM International Conference on AI in Finance (ICAIF)*, 2022.
5. A. Angiuli, J.-P. Fouque, R. Hu, A. Raydan. **Deep Reinforcement Learning for Infinite Horizon Mean Field Problems in Continuous Spaces**. *Journal of Machine Learning*, Vol. 4(1), 2025.
6. A. Angiuli, J.-P. Fouque, M. Laurière, M. Zhang. **Convergence of Multi-Scale Reinforcement Q-Learning Algorithms for Mean Field Game and Control Problems**. arXiv:2312.06659, 2023.
7. A. Angiuli, J.-P. Fouque, M. Laurière, M. Zhang. **Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games**. *Applied Mathematics & Optimization*, 2026.