Reinforcement Learning Algorithm for Mixed Mean Field Control Games

Jean-Pierre Fouque

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

Collaborators:
Andrea Angiuli, Nils Detering, Mathieu Laurière, and Jimin Lin

https://arxiv.org/abs/2205.02330
Journal of Machine Learning 2023

Stochastic Control & Financial Engineering Conference
Princeton University, NJ
June 23, 2023
Unified Reinforcement Q-Learning for Mean Field Game and Control Problems
Andrea Angiuli, JPF, and Mathieu Laurière

Mathematics of Control, Signals, and Systems (MCSS), 2022

https://arxiv.org/abs/2006.13912
Find \((\hat{\alpha}, \hat{\mu})\) such that the following two conditions hold:

\begin{enumerate}
\item \(\hat{\alpha}\) is the minimizer of

\[
\alpha \mapsto J_{AMFG}^{\alpha, \hat{\mu}}(\alpha; \hat{\mu}) = \mathbb{E} \left[ \sum_{n=0}^{+\infty} \gamma^n f(X_{n+1}^{\alpha, \hat{\mu}}, \alpha(X_n^{\alpha, \hat{\mu}}), \hat{\mu}) \right],
\]

where the process \(X^{\alpha, \hat{\mu}}\) follows the transitions:

\[
X_{n+1}^{\alpha, \hat{\mu}} \sim p(\cdot | X_n^{\alpha, \hat{\mu}}, \alpha(X_n^{\alpha, \hat{\mu}}), \hat{\mu})
\]

with initial distribution \(X_0^{\alpha, \hat{\mu}} \sim \mu_0\);

\item Asymptotic fixed point: \(\hat{\mu} = \lim_{n \to +\infty} \mathcal{L}(X_n^{\hat{\alpha}, \hat{\mu}})\).
\end{enumerate}

The control does not depend on time (\(f\) and \(p\) do not depend on time).

Intuitively, it corresponds to the situation in which an infinitesimal player wants to join a crowd of players who are already in the asymptotic regime (as time goes to infinity). This stationary distribution is a Nash equilibrium if the new player joining the crowd has no interest in deviating from this asymptotic behavior.
MFG Asymptotic formulation

Find \((\hat{\alpha}, \hat{\mu})\) such that the following two conditions hold:

1. \(\hat{\alpha}\) is the minimizer of

\[
\alpha \mapsto J^{AMFG}(\alpha; \hat{\mu}) = \mathbb{E} \left[ \sum_{n=0}^{+\infty} \gamma^n f(X_{n+1}^{\alpha, \hat{\mu}}, \alpha(X_n^{\alpha, \hat{\mu}}), \hat{\mu}) \right],
\]

where the process \(X_{\alpha, \hat{\mu}}\) follows the transitions:

\[
X_{n+1}^{\alpha, \hat{\mu}} \sim p(\cdot | X_{n}^{\alpha, \hat{\mu}}, \alpha(X_n^{\alpha, \hat{\mu}}), \hat{\mu})
\]

with initial distribution \(X_{0}^{\alpha, \hat{\mu}} \sim \mu_0\);

2. Asymptotic fixed point: \(\hat{\mu} = \lim_{n \to +\infty} \mathcal{L}(X_{n}^{\hat{\alpha}, \hat{\mu}})\).

The control does not depend on time \((f\text{ and } p\text{ do not depend on time})\).

Intuitively, it corresponds to the situation in which an infinitesimal player wants to join a crowd of players who are already in the asymptotic regime (as time goes to infinity). This stationary distribution is a Nash equilibrium if the new player joining the crowd has no interest in deviating from this asymptotic behavior.
MFG Asymptotic formulation

Find \((\hat{\alpha}, \hat{\mu})\) such that the following two conditions hold:

1. **\(\hat{\alpha}\) is the minimizer of**

\[
\alpha \mapsto J^{AMFG}(\alpha; \hat{\mu}) = \mathbb{E} \left[ \sum_{n=0}^{+\infty} \gamma^n f(X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu}) \right],
\]

where the process \(X^{\alpha,\hat{\mu}}\) follows the transitions:

\[
X_{n+1}^{\alpha,\hat{\mu}} \sim p \left( \cdot | X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu} \right)
\]

with initial distribution \(X_0^{\alpha,\hat{\mu}} \sim \mu_0\);

2. **Asymptotic fixed point:** \(\hat{\mu} = \lim_{n \to +\infty} \mathcal{L}(X_n^{\hat{\alpha},\hat{\mu}})\).

The control does not depend on time (\(f\) and \(p\) do not depend on time).

Intuitively, it corresponds to the situation in which an infinitesimal player wants to join a crowd of players who are already in the asymptotic regime (as time goes to infinity). This stationary distribution is a **Nash equilibrium** if the new player joining the crowd has no interest in deviating from this asymptotic behavior.
Find \( \alpha^* \) such that \( \alpha^* \) is the minimizer of

\[
\alpha \mapsto J^{AMFC}(\alpha) = \mathbb{E} \left[ \sum_{n=0}^{+\infty} \gamma^n f(X_{n}^{\alpha}, \alpha(X_{n}^{\alpha}), \mu^{\alpha}) \right],
\]

where the process \( X^\alpha \) follows the transitions

\[
X_{n+1}^{\alpha} \sim p(\cdot | X_{n}^{\alpha}, \alpha(X_{n}^{\alpha}), \mu^{\alpha})
\]

with initial distribution \( X_0^{\alpha} \sim \mu_0 \), and with the notation \( \mu^{\alpha} = \lim_{n \to +\infty} \mathcal{L}(X_{n}^{\alpha}) \).

Here too, the control is independent of time.

Intuitively, this problem can be viewed as the one posed to a central planner who wants to find the optimal stationary distribution such that the cost for the society is minimal when a new agent joins the crowd.
Find $\alpha^*$ such that $\alpha^*$ is the minimizer of

$$
\alpha \mapsto J^{AMFC}(\alpha) = \mathbb{E}\left[ \sum_{n=0}^{+\infty} \gamma^n f(X_n^\alpha, \alpha(X_n^\alpha), \mu^\alpha) \right],
$$

where the process $X^\alpha$ follows the transitions

$$
X_{n+1}^\alpha \sim p(\cdot | X_n^\alpha, \alpha(X_n^\alpha), \mu^\alpha)
$$

with initial distribution $X_0^\alpha \sim \mu_0$, and with the notation

$$
\mu^\alpha = \lim_{n \to +\infty} \mathcal{L}(X_n^\alpha).
$$

Here too, the control is independent of time.

Intuitively, this problem can be viewed as the one posed to a central planner who wants to find the optimal stationary distribution such that the cost for the society is minimal when a new agent joins the crowd.
Find $\alpha^*$ such that $\alpha^*$ is the minimizer of

$$\alpha \mapsto J^{AMFC}(\alpha) = \mathbb{E} \left[ \sum_{n=0}^{+\infty} \gamma^n f(X_\alpha^n, \alpha(X_\alpha^n), \mu_\alpha) \right],$$

where the process $X_\alpha$ follows the transitions

$$X_\alpha^{n+1} \sim p(\cdot|X_\alpha^n, \alpha(X_\alpha^n), \mu_\alpha)$$

with initial distribution $X_0^\alpha \sim \mu_0$, and with the notation

$$\mu_\alpha = \lim_{n \to +\infty} \mathcal{L}(X_\alpha^n).$$

Here too, the control is independent of time.

Intuitively, this problem can be viewed as the one posed to a central planner who wants to find the optimal stationary distribution such that the cost for the society is minimal when a new agent joins the crowd.
Although the AMFG and AMFC problems are defined using an initial distribution $\mu_0$, one expects that under suitable conditions, \textit{ergodicity} in particular, the optimal control is independent of this initial distribution.

We also consider the stationary problems:

**Stationary MFG problem**: same as AMFG except that $\mu_0 = \hat{\mu}$ is part of the problem.

**Stationary MFC problem**: same as AMFC except that $\mu_0 = \mu^\alpha$ is part of the problem.

Connecting the two formulations: we have:

$$\hat{\alpha}^{AMFG}(x) = \hat{\alpha}^{SMFG}(x), \quad \forall x$$

and

$$\alpha^{*AMFC}(x) = \alpha^{*SMFC}(x), \quad \forall x$$
Although the AMFG and AMFC problems are defined using an initial distribution $\mu_0$, one expects that under suitable conditions, *ergodicity* in particular, the optimal control is independent of this initial distribution. We also consider the stationary problems:

**Stationary MFG problem**: same as AMFG except that $\mu_0 = \hat{\mu}$ is part of the problem.

**Stationary MFC problem**: same as AMFC except that $\mu_0 = \mu^\alpha$ is part of the problem.

Connecting the two formulations: we have:

$$\hat{\alpha}^{AMFG}(x) = \hat{\alpha}^{SMFG}(x), \quad \forall x$$

and

$$\alpha^{*AMFC}(x) = \alpha^{*SMFC}(x), \quad \forall x$$
Although the AMFG and AMFC problems are defined using an initial distribution $\mu_0$, one expects that under suitable conditions, ergodicity in particular, the optimal control is independent of this initial distribution. We also consider the stationary problems:

**Stationary MFG problem**: same as AMFG except that $\mu_0 = \hat{\mu}$ is part of the problem.

**Stationary MFC problem**: same as AMFC except that $\mu_0 = \mu^\alpha$ is part of the problem.

Connecting the two formulations: we have:

$$\hat{\alpha}^{AMFG}(x) = \hat{\alpha}^{SMFG}(x), \quad \forall x$$

and

$$\alpha^{*AMFC}(x) = \alpha^{*SMFC}(x), \quad \forall x$$
Although the AMFG and AMFC problems are defined using an initial distribution $\mu_0$, one expects that under suitable conditions, ergodicity in particular, the optimal control is independent of this initial distribution. We also consider the stationary problems:

**Stationary MFG problem**: same as AMFG except that $\mu_0 = \hat{\mu}$ is part of the problem.

**Stationary MFC problem**: same as AMFC except that $\mu_0 = \mu^\alpha$ is part of the problem.

**Connecting the two formulations**: we have:

$$\hat{\alpha}^{AMFG}(x) = \hat{\alpha}^{SMFG}(x), \quad \forall x$$

and

$$\alpha^{*AMFC}(x) = \alpha^{*SMFC}(x), \quad \forall x$$
In the context of Asymptotic MFG, we can view the problem faced by an infinitesimal agent among the crowd as an MDP parameterized by the population distribution.

Then, the optimal $Q$-function is defined, for a given $\mu$, by

$$Q^*_\mu(x, a) = \min_{\alpha} \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_{tn}, \alpha(X_{tn}), \mu) \mid X_{t0} = x, A_{t0} = a \right]$$

where the cost function $f(x, a, \mu)$ depends on the fixed $\mu$ as well as the transition probabilities $p(x'|x, a, \mu)$.

Since $\mu$ is fixed, one obtains the classical Bellman equation:

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

This function characterizes the optimal cost-to-go for an agent starting at state $x$, using action $a$ for the first step, and then acting optimally for the rest of the time steps, while the population distribution is given by $\mu$ (for every time step).
In the context of Asymptotic MFG, we can view the problem faced by an infinitesimal agent among the crowd as an MDP parameterized by the population distribution. Then, the optimal $Q$-function is defined, for a given $\mu$, by

$$Q_\mu^*(x, a) = \min_{\alpha} \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_{tn}, \alpha(X_{tn}), \mu) \middle| X_{t0} = x, A_{t0} = a \right]$$

where the cost function $f(x, a, \mu)$ depends on the fixed $\mu$ as well as the transition probabilities $p(x'|x, a, \mu)$

Since $\mu$ is fixed, one obtains the classical Bellman equation:

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

This function characterizes the optimal cost-to-go for an agent starting at state $x$, using action $a$ for the first step, and then acting optimally for the rest of the time steps, while the population distribution is given by $\mu$ (for every time step).
In the context of Asymptotic MFG, we can view the problem faced by an infinitesimal agent among the crowd as an MDP parameterized by the population distribution.

Then, the optimal $Q$-function is defined, for a given $\mu$, by

$$Q^*_\mu(x, a) = \min_{\alpha} \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_{tn}, \alpha(X_{tn}), \mu) \mid X_{t0} = x, A_{t0} = a \right]$$

where the cost function $f(x, a, \mu)$ depends on the fixed $\mu$ as well as the transition probabilities $p(x' \mid x, a, \mu)$.

Since $\mu$ is fixed, one obtains the classical Bellman equation:

$$Q^*_\mu(x, a) = f(x, a, \mu) + \gamma \sum_{x' \in \mathcal{X}} p(x' \mid x, a, \mu) \min_{a'} Q^*_\mu(x', a'), \quad (x, a) \in \mathcal{X} \times \mathcal{A}$$

This function characterizes the optimal cost-to-go for an agent starting at state $x$, using action $a$ for the first step, and then acting optimally for the rest of the time steps, while the population distribution is given by $\mu$ (for every time step).
For the two rates of update $\rho_k^\mu$ and $\rho_k^Q$:  
\[
\begin{align*}
\mu_{k+1} &= \mu_k + \rho_k^\mu P(Q_k, \mu_k), \\
Q_{k+1} &= Q_k + \rho_k^Q T(Q_k, \mu_k),
\end{align*}
\]

with
\[
\begin{align*}
P(Q, \mu)(x) &= (\mu P^Q\mu)(x) - \mu(x), \quad x \in X, \\
T(Q, \mu)(x, a) &= f(x, a, \mu) + \gamma \sum_{x'} p(x'|x, a, \mu) \min_{a'} Q(x', a') - Q(x, a),
\end{align*}
\]
and
\[
(\mu P^Q\mu)(x) = \sum_{x_0} \mu(x_0) P^Q\mu(x_0, x),
\]

where
\[
P^Q\mu(x, x') = p(x'|x, \arg\min_a Q(x, a), \mu),
\]
is the transition matrix when the population distribution is $\mu$ and the agent uses the optimal control according to $Q$. 

Jean-Pierre Fouque (UCSB)
Two-timescale approach for the Asymptotic MFG

If \( \rho_k^\mu \ll \rho_k^Q \), set \( \rho_k^\mu / \rho_k^Q = \epsilon \ll 1 \), then, in “continuous time”:

\[
\begin{align*}
\dot{\mu}_t &= \mathcal{P}(Q_t, \mu_t) \\
\dot{Q}_t &= \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t)
\end{align*}
\]

For any fixed \( \mu \), as \( \epsilon \to 0 \), the solution of the second ODE is expected to converge to an equilibrium \( Q_\mu \) such that \( \mathcal{T}(Q_\mu, \mu) = 0 \), meaning that \( Q \) is the value function of an infinitesimal agent facing the crowd distribution \( \mu \).

Then, the first ODE becomes

\[\dot{\mu}_t = \mathcal{P}(Q_{\mu t}, \mu_t)\]

which is expected to converge as \( t \to +\infty \) to \( \mu_\infty \) satisfying

\[\mathcal{P}(Q_{\mu_\infty}, \mu_\infty) = 0\]

This condition means that \( \mu_\infty \) and the associated control given by \( \hat{\alpha}(x) = \arg\min_a Q_{\mu_\infty}(x, a) \) form a Nash equilibrium.
Two-timescale approach for the Asymptotic MFG

If $\rho_k^\mu \ll \rho_k^Q$, set $\rho_k^\mu / \rho_k^Q = \epsilon \ll 1$, then, in “continuous time”:

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

For any fixed $\mu$, as $\epsilon \to 0$, the solution of the second ODE is expected to converge to an equilibrium $Q_\mu$ such that $\mathcal{T}(Q_\mu, \mu) = 0$, meaning that $Q$ is the value function of an infinitesimal agent facing the crowd distribution $\mu$.

Then, the first ODE becomes

$$
\dot{\mu}_t = \mathcal{P}(Q_\mu, \mu_t)
$$

which is expected to converge as $t \to +\infty$ to $\mu_\infty$ satisfying

$$
\mathcal{P}(Q_\infty, \mu_\infty) = 0
$$

This condition means that $\mu_\infty$ and the associated control given by $\hat{\alpha}(x) = \arg\min_a Q_{\mu_\infty}(x, a)$ form a Nash equilibrium.
Two-timescale approach for the Asymptotic MFG

If $\rho_k^{\mu} \ll \rho_k^Q$, set $\rho_k^{\mu}/\rho_k^Q = \epsilon \ll 1$, then, in “continuous time”:

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

For any fixed $\mu$, as $\epsilon \to 0$, the solution of the second ODE is expected to converge to an equilibrium $Q_\mu$ such that $\mathcal{T}(Q_\mu, \mu) = 0$, meaning that $Q$ is the value function of an infinitesimal agent facing the crowd distribution $\mu$. Then, the first ODE becomes

$$
\dot{\mu}_t = \mathcal{P}(Q_{\mu_t}, \mu_t)
$$

which is expected to converge as $t \to +\infty$ to $\mu_\infty$ satisfying

$$
\mathcal{P}(Q_{\mu_\infty}, \mu_\infty) = 0
$$

This condition means that $\mu_\infty$ and the associated control given by $\hat{\alpha}(x) = \arg \min_a Q_{\mu_\infty}(x, a)$ form a Nash equilibrium.
If $\rho_k^\mu \ll \rho_k^Q$, set $\rho_k^\mu / \rho_k^Q = \epsilon \ll 1$, then, in “continuous time”:

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

For any fixed $\mu$, as $\epsilon \to 0$, the solution of the second ODE is expected to converge to an equilibrium $Q_\mu$ such that $\mathcal{T}(Q_\mu, \mu) = 0$, meaning that $Q$ is the value function of an infinitesimal agent facing the crowd distribution $\mu$. Then, the first ODE becomes

$$
\dot{\mu}_t = \mathcal{P}(Q_\mu, \mu_t)
$$

which is expected to converge as $t \to +\infty$ to $\mu_\infty$ satisfying

$$
\mathcal{P}(Q_\mu, \mu_\infty) = 0
$$

This condition means that $\mu_\infty$ and the associated control given by $\hat{\alpha}(x) = \arg\min_a Q_{\mu_\infty}(x, a)$ form a **Nash equilibrium.**
For a control $\alpha$ and $(x, a)$, we define the control

$$\tilde{\alpha}(x') = \begin{cases} 
    a & \text{if } x' = x \\
    \alpha(x) & \text{for } x' \neq x
\end{cases}$$

Our modified $Q$-function is given by

$$Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \bigg| X_0 = x, A_0 = a \right].$$

We then obtain that the optimal $Q^*(x, a) = \min_{\alpha} Q^\alpha(x, a)$ satisfies the modified MKV Bellman equation

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

where the optimal control $\alpha^*$ is given by $\alpha^*(x) = \arg \min_{a} Q^*(x, a)$, the control $\tilde{\alpha}^*$ is defined for $x$ and $a$, and $\mu^* := \mu^{\tilde{\alpha}^*}$.

The optimal value function is $V^*(x) = \min_{a} Q^*(x, a)$. 
Action-value function for the Asymptotic MFC

For a control $\alpha$ and $(x, a)$, we define the control

$$\tilde{\alpha}(x') = \begin{cases} a & \text{if } x' = x \\ \alpha(x) & \text{for } x' \neq x \end{cases}$$

Our modified $Q$-function is given by

$$Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^{\alpha}) \bigg| X_0 = x, A_0 = a \right].$$

We then obtain that the optimal $Q^*(x, a) = \min_\alpha Q^\alpha(x, a)$ satisfies the modified MKV Bellman equation

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

where the optimal control $\alpha^*$ is given by $\alpha^*(x) = \arg \min_a Q^*(x, a)$, the control $\tilde{\alpha}^*$ is defined for $x$ and $a$, and $\tilde{\mu}^* := \mu^{\alpha^*}$.

The optimal value function is $V^*(x) = \min_a Q^*(x, a)$. 
For a control $\alpha$ and $(x, a)$, we define the control

$$\tilde{\alpha}(x') = \begin{cases} 
  a & \text{if } x' = x \\
  \alpha(x) & \text{for } x' \neq x 
\end{cases}$$

Our modified $Q$-function is given by

$$Q^\alpha(x, a) = f(x, a, \mu \tilde{\alpha}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \Bigg| X_0 = x, A_0 = a \right].$$

We then obtain that the optimal $Q^*(x, a) = \min_\alpha Q^\alpha(x, a)$ satisfies the modified MKV Bellman equation

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

where the optimal control $\alpha^*$ is given by $\alpha^*(x) = \arg \min_a Q^*(x, a)$, the control $\tilde{\alpha}^*$ is defined for $x$ and $a$, and $\tilde{\mu}^* := \mu^{\tilde{\alpha}^*}$.

The optimal value function is $V^*(x) = \min_a Q^*(x, a)$. 
Action-value function for the Asymptotic MFC

For a control \( \alpha \) and \((x, a)\), we define the control

\[
\tilde{\alpha}(x') = \begin{cases} 
  a & \text{if } x' = x \\
  \alpha(x) & \text{for } x' \neq x 
\end{cases}
\]

Our modified \( Q \)-function is given by

\[
Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E}\left[ \sum_{n=1}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu^\alpha) \bigg| X_0 = x, A_0 = a \right].
\]

We then obtain that the optimal \( Q^*(x, a) = \min_\alpha Q^\alpha(x, a) \) satisfies the modified MKV Bellman equation

\[
Q^*(x, a) = f(x, a, \mu^*) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a, \mu^*) \min_{a'} Q^*(x', a'), \quad (x, a) \in \mathcal{X} \times \mathcal{A}
\]

where the optimal control \( \alpha^* \) is given by \( \alpha^*(x) = \arg \min_a Q^*(x, a) \), the control \( \tilde{\alpha}^* \) is defined for \( x \) and \( a \), and \( \mu^* := \mu^{\tilde{\alpha}^*} \).

The optimal value function is \( V^*(x) = \min_a Q^*(x, a) \).
Recall the classical MFG $Q$-function:

$$Q_{\mu}^{\alpha}(x, a) = f(x, a, \mu) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{t_n}, \alpha(X_{t_n}), \mu) \bigg| X_{t_0} = x, A_{t_0} = a \right]$$

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

and the modified MFC $Q$-function:

$$Q^{\alpha}(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{t_n}, \alpha(X_{t_n}), \mu^{\alpha}) \bigg| X_{t_0} = x, A_{t_0} = a \right]$$

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

We have:

$$Q^{\alpha} = Q_{\mu}^{\alpha} + \Delta_{\mu} f + \gamma \Delta_{\mu} p \left( Q_{\mu}^{\alpha} \right)$$

with the additional “derivatives” with respect to $\mu$ involved in MFC.
Recall the classical MFG $Q$-function:

\[
Q^\alpha_\mu(x, a) = f(x, a, \mu) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{tn}, \alpha(X_{tn}), \mu) \right| X_{t_0} = x, A_{t_0} = a \right] \\
= f(x, a, \mu) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a, \mu) Q^\alpha_\mu(x', \alpha(x'))
\]

and the modified MFC $Q$-function:

\[
Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{tn}, \alpha(X_{tn}), \mu^{\tilde{\alpha}}) \right| X_{t_0} = x, A_{t_0} = a \right] \\
= f(x, a, \mu^{\tilde{\alpha}}) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a, \mu^{\tilde{\alpha}}) Q^\alpha(x', \alpha(x'))
\]

We have:

\[
Q^\alpha = Q^\alpha_\mu + \Delta_\mu f + \gamma \Delta_\mu p (Q^\alpha_\mu)
\]

with the additional "derivatives" with respect to $\mu$ involved in MFC.
Recall the classical MFG $Q$-function:

$$Q_\mu^\alpha(x, a) = f(x, a, \mu) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{t_n}, \alpha(X_{t_n}), \mu) \bigg| X_{t_0} = x, A_{t_0} = a \right]$$

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

and the modified MFC $Q$-function:

$$Q^\alpha(x, a) = f(x, a, \mu^{\tilde{\alpha}}) + \mathbb{E} \left[ \sum_{n=1}^{\infty} \gamma^n f(X_{t_n}, \alpha(X_{t_n}), \mu^{\alpha}) \bigg| X_{t_0} = x, A_{t_0} = a \right]$$

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

We have:

$$Q^\alpha = Q_\mu^\alpha + \Delta_\mu f + \gamma \Delta_\mu p (Q_\mu^\alpha)$$

with the additional “derivatives” with respect to $\mu$ involved in MFC.
Two-timescale approach for MFC

If $\rho_k^\mu \gg \rho_k^Q$, set $\rho_k^Q / \rho_k^\mu = \epsilon \ll 1$, then, in “continuous time”:

$$
\begin{align*}
\dot{\mu}_t &= \frac{1}{\epsilon} P(Q_t, \mu_t), \\
\dot{Q}_t &= T(Q_t, \mu_t),
\end{align*}
$$

Here, for any fixed $Q$, the solution of the first ODE is expected to converge as $\epsilon \to 0$ to an equilibrium $\mu_Q$ such that $P(Q, \mu_Q) = 0$, meaning that $\mu_Q$ is the asymptotic distribution of a population in which every agent uses the control $\alpha(x) = \arg \min_a Q(x, a)$.

Then the second ODE becomes

$$
\dot{Q}_t(x, a) = T(Q_t(x, a), \tilde{\mu}_{Q_t}),
$$

where $\tilde{\mu}_{Q_t}$ is define at $(x, a)$ for the control $\alpha(\cdot) = \arg \min_{a'} Q_t(\cdot, a')$. This is consistent with the update of $Q$ and with the algorithm which follows.
If $\rho^\mu_k \gg \rho^Q_k$, set $\frac{\rho^Q_k}{\rho^\mu_k} = \epsilon \ll 1$, then, in “continuous time”:

\[
\begin{align*}
\dot{\mu}_t &= \frac{1}{\epsilon} P(Q_t, \mu_t), \\
\dot{Q}_t &= T(Q_t, \mu_t),
\end{align*}
\]

Here, for any fixed $Q$, the solution of the first ODE is expected to converge as $\epsilon \to 0$ to an equilibrium $\mu_Q$ such that $P(Q, \mu_Q) = 0$, meaning that $\mu_Q$ is the asymptotic distribution of a population in which every agent uses the control $\alpha(x) = \arg \min_a Q(x, a)$.

Then the second ODE becomes

\[
\dot{Q}_t(x, a) = T(Q_t(x, a), \tilde{\mu}_Q t),
\]

where $\tilde{\mu}_Q t$ is define at $(x, a)$ for the control $\alpha(\cdot) = \arg \min_{a'} Q_t(\cdot, a')$. This is consistent with the update of $Q$ and with the algorithm which follows.
Two-timescale approach for MFC

If $\rho^\mu_k \gg \rho^Q_k$, set $\rho^Q_k / \rho^\mu_k = \epsilon \ll 1$, then, in "continuous time":

$$
\begin{align*}
\dot{\mu}_t &= \frac{1}{\epsilon} P(Q_t, \mu_t), \\
\dot{Q}_t &= T(Q_t, \mu_t),
\end{align*}
$$

Here, for any fixed $Q$, the solution of the first ODE is expected to converge as $\epsilon \to 0$ to an equilibrium $\mu_Q$ such that $P(Q, \mu_Q) = 0$, meaning that $\mu_Q$ is the asymptotic distribution of a population in which every agent uses the control $\alpha(x) = \arg \min_a Q(x, a)$.

Then the second ODE becomes

$$
\dot{Q}_t(x, a) = T(Q_t(x, a), \tilde{\mu}_Q t),
$$

where $\tilde{\mu}_Q$ is define at $(x, a)$ for the control $\alpha(\cdot) = \arg \min_{a'} Q_t(\cdot, a')$. This is consistent with the update of $Q$ and with the algorithm which follows.
Two-timescale approach for MFC

We have

\[ \dot{Q}_t(x, a) = T(Q_t(x, a), \tilde{\mu}_{Q_t}) \]

The limit \( Q_\infty \) as \( t \to +\infty \) satisfies

\[ T(Q_\infty, \tilde{\mu}_{Q_\infty}) = 0 \]

meaning that \( Q_\infty = Q^* \) satisfies the MFC Bellman equation.

In fact, only true along \( a = \alpha^*(x) \), or equivalently, at the level of the value functions.

Therefore the control

\[ \alpha^*(x) = \arg \min_a Q_\infty(x, a) \]

is an MFC optimum for the asymptotic formulation and the induced optimal distribution is \( \mu_{Q_\infty} \).
Two-timescale approach for MFC

We have

\[ \dot{Q}_t(x, a) = T(Q_t(x, a), \tilde{\mu}_{Q_t}) \]

The limit \( Q_\infty \) as \( t \to +\infty \) satisfies

\[ T(Q_\infty, \tilde{\mu}_{Q_\infty}) = 0 \]

meaning that \( Q_\infty = Q^* \) satisfies the MFC Bellman equation.

In fact, only true along \( a = \alpha^*(x) \), or equivalently, at the level of the value functions.

Therefore the control

\[ \alpha^*(x) = \arg\min_a Q_\infty(x, a) \]

is an MFC optimum for the asymptotic formulation and the induced optimal distribution is \( \mu_{Q_\infty} \).
**MDP with Mean Field interaction**: Interaction of the representative agent with the environment. When the current state of the environment is $X_{tn}$, given a distribution $\mu_{tn}$ and an action $A_{tn}$, the environment produces the new state $X_{tn+1}$ and incurs a cost $f_{tn+1}$ calculated by starting from the current state of the environment $X_{tn}$ and using the transition controlled by $A_{tn}$ and parameterized by $\mu_{tn}$. 
Algorithm 1 Unified Two Timescales Mean Field Q-learning - Tabular version

**Require:** \( \tau = \{ t_0 = 0, \ldots, t_{|\tau| - 1} = T \} \) with \( t_0 = 0 < \cdots < t_{|\tau| - 1} = T \) : time steps,
\( \mathcal{X} = \{ x_0, \ldots, x_{|\mathcal{X}| - 1} \} \) : finite state space,
\( \mathcal{A} = \{ a_0, \ldots, a_{|\mathcal{A}| - 1} \} \) : finite action space,
\( \mu_0 \) : initial distribution of the representative player,
\( \{ \epsilon_n \}_{n=0}^{\tau-1} \) : time-decaying factor related to the \( \epsilon \)-greedy policy,
tol\( _\mu \), tol\( _Q \) : break rule tolerances.

1: **Initialization:** \( Q^0(x, a) = 0 \) for all \( (x, a) \in \mathcal{X} \times \mathcal{A} \), \( \mu^0_n = \left[ \frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|} \right] \) for \( n = 0, \ldots, |\tau| - 1 \)
2: for each episode \( k = 1, 2, \ldots \) do
3: **Initialization:** Sample \( X_{t_0} \sim \mu^k_{t_0} \) and set \( Q^k = Q^{k-1} \)
4: for \( n \leftarrow 0 \) to \( |\tau| - 1 \) do
5: **Update \( \mu \):**
   \( \mu^k_{t_n} = \mu^k_{t_{n-1}} + \rho^k_{\mu} (\delta(X_{t_n}) - \mu^k_{t_{n-1}}) \)
   where \( \delta(X_{t_n}) = \left[ \mathbf{1}_{x_0}(X_{t_n}), \ldots, \mathbf{1}_{x_{|\mathcal{X}|-1}}(X_{t_n}) \right] \)
6: **Choose action** \( A_{t_n} \) using the \( \epsilon_n \)-greedy policy derived from \( Q^k(X_{t_n}, \cdot) \)
7: **Observe cost** \( f_{t_{n+1}} = f(X_{t_n}, A_{t_n}, \mu^k_{t_n}) \) and state \( X_{t_{n+1}} \) provided by the environment
8: **Update \( Q \):**
   \( Q^k(X_{t_n}, A_{t_n}) = Q^k(X_{t_n}, A_{t_n}) + \rho^Q_{k,n,X_{t_n},A_{t_n}} [B - Q^k(X_{t_n}, A_{t_n})] \)
   where \( B := f_{t_{n+1}} + \gamma \min_{a' \in \mathcal{A}} Q^k(X_{t_{n+1}}, a') \)
9: if \( ||\mu^k - \mu^{k-1}|| \leq tol_\mu \) and \( ||Q^k - Q^{k-1}|| < tol_Q \) then
10: break
11: end if
12: end for
The learning rates are:

\[ \rho_{k,n,x,a}^Q = \frac{1}{(1 + \#|(x,a,k,n)|)\omega^Q}, \quad \rho_k^\mu = \frac{1}{(1 + k)\omega^\mu}, \]

where \#|(x,a,k,n)| is the number of times that the algorithm visited state \(x\) and performed action \(a\) until episode \(k\) and time \(t_n\).

The exponent \(\omega^Q\) can take values in \((\frac{1}{2}, 1)\).

The value of \(\omega^\mu\) is chosen depending on the value of \(\omega^Q\) and the MFG (competitive) or MFC (cooperative) nature of the problem we want to solve.

The algorithm is run over 80K episodes over the interval \([0, T]\).
Learning rates and two timescales

**MFG**: learning rates over $80 \times 10^3$ episodes, $(\omega^Q, \omega^\mu) = (0.55, 0.85)$:
\[ \rho^\mu \ll \rho^Q \]

**MFC**: learning rates over $80 \times 10^3$ episodes, $(\omega^Q, \omega^\mu) = (0.65, 0.15)$
\[ \rho^\mu \gg \rho^Q \]
$dX_t = \alpha(X_t)dt + \sigma dW_t$

The drift is $b(x, \alpha, \mu) = \alpha(x)$, with $d = 1$, and the mean-field interactions are through the first moment $m = \int x\mu(x)dx$ in the cost function:

$$f(x, \alpha, \mu) = \frac{1}{2}\alpha^2 + c_1 (x - c_2 m)^2 + c_3 (x - c_4)^2 + c_5 m^2$$

Here the parameters $c_1, c_2, c_3, c_5 \in \mathbb{R}_+$, $c_2 \leq 1$, and $c_4 \in \mathbb{R}$ are constant.

- $\frac{1}{2}\alpha^2$: quadratic cost for controlling the drift,
- $c_1 (x - c_2 m)^2$: mean field interaction ($c_2 = 1$: mean-reverting effect),
- $c_3 (x - c_4)^2$: target position $c_4$,
- $c_5 m^2$: penalizes the population when its mean $m$ is far away from zero.
Benchmark model

\[ dX_t = \alpha(X_t)dt + \sigma dW_t \]

The drift is \( b(x, \alpha, \mu) = \alpha(x) \), with \( d = 1 \), and the mean-field interactions are through the first moment \( m = \int x\mu(x)dx \) in the cost function:

\[
f(x, \alpha, \mu) = \frac{1}{2} \alpha^2 + c_1 (x - c_2 m)^2 + c_3 (x - c_4)^2 + c_5 m^2
\]

Here the parameters \( c_1, c_2, c_3, c_5 \in \mathbb{R}^+ \), \( c_2 \leq 1 \), and \( c_4 \in \mathbb{R} \) are constant.

- \( \frac{1}{2} \alpha^2 \): quadratic cost for controlling the drift,
- \( c_1 (x - c_2 m)^2 \): mean field interaction (\( c_2 = 1 \): mean-reverting effect),
- \( c_3 (x - c_4)^2 \): target position \( c_4 \),
- \( c_5 m^2 \): penalizes the population when its mean \( m \) is far away from zero.
Controls (averaged over 10 runs) and Equilibrium distributions (averaged over last 10K episodes)

MFG: \((\omega^Q, \omega^\mu) = (0.55, 0.85)\)  
MFC: \((\omega^Q, \omega^\mu) = (0.65, 0.15)\)
Reinforcement Learning for Mean Field Games, with Applications to Economics
Andrea Angiuli, JPF, and Mathieu Laurière

To appear in the Handbook of Machine Learning for Financial Markets
Cambridge University Press
Editors: Agostino Capponi and Charles-Albert Lehalle
A Mean Field Control Game (MFCG) can be interpreted as a competitive game between collaborating groups and its solution as a Nash equilibrium between the groups. Within each group the players coordinate their strategies. In the limit of a large number of large groups, we will use

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

so that the distribution $\tilde{\mu}$ of the agent’s group is updated faster than the Q-function (MFC), which is itself update faster than the distribution $\mu$ of the overall population (MFG).
A Mean Field Control Game (MFCG) can be interpreted as a competitive game between collaborating groups and its solution as a Nash equilibrium between the groups. Within each group the players coordinate their strategies.

In the limit of a large number of large groups, we will use

\[ \rho^\mu \ll \rho^Q \ll \rho^{\tilde{\mu}} \]

so that the distribution \( \tilde{\mu} \) of the agent’s group is updated faster than the Q-function (MFC), which is itself update faster than the distribution \( \mu \) of the overall population (MFG).
Idea: our algorithm can handle three scales!

Reinforcement Learning Algorithm for Mixed Mean Field Control Games

by Andrea Angiuli, Nils Detering, JPF, and Jimin Lin

https://arxiv.org/abs/2205.02330

A Mean Field Control Game (MFCG) can be interpreted as a competitive game between collaborating groups and its solution as a Nash equilibrium between the groups. Within each group the players coordinate their strategies. In the limit of a large number of large groups, we will use

\[ \rho^\mu \ll \rho^Q \ll \rho^{\tilde{\mu}} \]

so that the distribution \( \tilde{\mu} \) of the agent’s group is updated faster than the Q-function (MFC), which is itself update faster than the distribution \( \mu \) of the overall population (MFG).
Infinite horizon MFCG on finite spaces

Given a cost function \( f \) defined on \( \mathcal{X} \times \mathcal{A} \times \Delta^{\mathcal{X}} \times \Delta^{\mathcal{X}} \) and a discount rate \( \gamma < 1 \), we now consider the following infinite horizon asymptotic mean field control game (MFCG) problem:

Find a strategy \( \hat{\alpha} \), and distribution \( \hat{\mu} \) such that:

1. **(best response)** \( \hat{\alpha} \) is the minimizer of

\[
J(\alpha; \hat{\mu}) = \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f \left( X_n^{\alpha,\hat{\mu}}, \alpha(X_n^{\alpha,\hat{\mu}}), \hat{\mu}, \mu^{\alpha} \right) \right],
\]

where \( X_0^{\alpha,\mu} \sim \mu_0 \) and

\[
\mathbb{P}(X_{n+1}^{\alpha,\hat{\mu}} = x' | X_n^{\alpha,\hat{\mu}} = x, \alpha(X_n^{\alpha,\hat{\mu}}) = a, \hat{\mu} = \mu, \mu^{\alpha} = \tilde{\mu}) = p(x' | x, a, \mu, \mu^{\alpha})
\]

and \( \mu^{\alpha} = \lim_{n \to \infty} \mathcal{L}(X_n^{\alpha,\hat{\mu}}) \).

2. **(fixed-point)** \( \hat{\mu} = \lim_{n \to \infty} \mathcal{L}(X_n^{\hat{\alpha},\hat{\mu}}) = \mu^{\hat{\alpha}} \).
Infinite horizon MFCG on finite spaces

Given a cost function $f$ defined on $\mathcal{X} \times \mathcal{A} \times \Delta^{\mathcal{X}} \times \Delta^{\mathcal{X}}$ and a discount rate $\gamma < 1$, we now consider the following infinite horizon asymptotic mean field control game (MFCG) problem:

Find a strategy $\hat{\alpha}$, and distribution $\hat{\mu}$ such that:

1. (best response) $\hat{\alpha}$ is the minimizer of

   $$J(\alpha; \hat{\mu}) = \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f \left( X_n^{\alpha, \hat{\mu}}, \alpha(X_n^{\alpha, \hat{\mu}}), \hat{\mu}, \mu^\alpha \right) \right],$$

   where $X_0^{\alpha, \mu} \sim \mu_0$ and

   $$\mathbb{P}(X_{n+1}^{\alpha, \hat{\mu}} = x' | X_n^{\alpha, \hat{\mu}} = x, \alpha(X_n^{\alpha, \hat{\mu}}) = a, \hat{\mu} = \mu, \mu^\alpha = \tilde{\mu}) = p(x' | x, a, \mu, \mu^\alpha).$$

   and $\mu^\alpha = \lim_{n \to \infty} \mathcal{L}(X_n^{\alpha, \hat{\mu}})$.

2. (fixed-point) $\hat{\mu} = \lim_{n \to \infty} \mathcal{L}(X_n^{\hat{\alpha}, \hat{\mu}}) = \mu^{\hat{\alpha}}$. 

Jean-Pierre Fouque (UCSB)
For an admissible control $\alpha : \mathcal{X} \to \mathcal{A}$, we define the new control $\alpha_{x,a}$ by

$$\alpha_{x,a}(x') = \begin{cases} a & \text{if } x' = x \\ \alpha(x) & \text{otherwise} \end{cases}$$

Given a global measure $\mu$ and strategy $\alpha$, the $Q$-function is given by:

$$Q_{\mu}^\alpha(x, a) = f(x, a, \mu, \mu^{\alpha_{x,a}}(x,a)) + \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu, \mu^{\alpha}(x,a)) | X_0 = x, A_0 = a \right]$$

One can then consider:

$$Q_{\mu}^*(x, a) := \min_{\alpha} Q_{\mu}^\alpha(x, a).$$

From the function $Q_{\mu}^*$ one obtains the control $\alpha^*(x) = \arg \min_a Q_{\mu}^*(x, a)$. The function $Q_{\mu}^*$ satisfies the Bellman equation:

$$Q_{\mu}^*(x, a) = f(x, a, \mu, \mu^{*}_{x,a}) + \gamma \sum_{x'} p(x' | x, a, \mu, \mu^{*}_{x,a}) \min_{a'} Q_{\mu}^*(x', a').$$
Q-matrix and MKV Bellman equation

For an admissible control $\alpha : \mathcal{X} \rightarrow \mathcal{A}$, we define the new control $\alpha_{x,a}$ by

$$\alpha_{x,a}(x') = \begin{cases} a & \text{if } x' = x \\ \alpha(x) & \text{otherwise} \end{cases}$$

Given a global measure $\mu$ and strategy $\alpha$, the $Q$-function is given by:

$$Q_\mu^\alpha(x, a) = f(x, a, \mu, \mu_{x,a}^\alpha) + \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu, \mu^\alpha)|X_0 = x, A_0 = a \right]$$

One can then consider:

$$Q_\mu^*(x, a) := \min_{\alpha} Q_\mu^\alpha(x, a).$$

From the function $Q_\mu^*$ one obtains the control $\alpha^*(x) = \arg \min_a Q_\mu^*(x, a)$. The function $Q_\mu^*$ satisfies the Bellman equation:

$$Q_\mu^*(x, a) = f(x, a, \mu, \mu_{x,a}^*) + \gamma \sum_{x'} \mathbb{P}(x'|x, a, \mu, \mu_{x,a}^*) \min_{a'} Q_\mu^*(x', a').$$
For an admissible control $\alpha : \mathcal{X} \to A$, we define the new control $\alpha_{x,a}$ by

$$
\alpha_{x,a}(x') = \begin{cases} 
    a & \text{if } x' = x \\
    \alpha(x) & \text{otherwise}
\end{cases}
$$

Given a global measure $\mu$ and strategy $\alpha$, the $Q$-function is given by:

$$
Q_\mu^\alpha(x, a) = f(x, a, \mu, \mu_{\alpha_{x,a}}) + \mathbb{E} \left[ \sum_{n=0}^{\infty} \gamma^n f(X_n, \alpha(X_n), \mu, \mu^\alpha) \middle| X_0 = x, A_0 = a \right]
$$

One can then consider:

$$
Q_\mu^*(x, a) := \min_{\alpha} Q_\mu^\alpha(x, a).
$$

From the function $Q_\mu^*$ one obtains the control $\alpha^*(x) = \text{arg min}_a Q_\mu^*(x, a)$. The function $Q_\mu^*$ satisfies the Bellman equation:

$$
Q_\mu^*(x, a) = f(x, a, \mu, \mu_{x,a}^*) + \gamma \sum_{x'} p(x'|x, a, \mu, \mu_{x,a}^*) \min_{a'} Q_\mu^*(x', a').
$$
Learning rates

With rates \( \rho_k^\mu < \rho_k^Q < \rho_k^{\mu^\alpha} \), \((\mu, \mu^\alpha, Q)\) are updated by

\[
\begin{align*}
\mu_{k+1} &= \mu_k + \rho_k^\mu \mathcal{P}(Q_k, \mu_k), \\
\mu_{k+1}^\alpha &= \mu_{k}^\alpha + \rho_k^{\mu^\alpha} \mathcal{P}(Q_k, \mu_k^\alpha), \\
Q_{k+1} &= Q_k + \rho_k^Q \mathcal{T}(Q_k, \mu_k, \mu_k^\alpha),
\end{align*}
\]

where

\[
\begin{align*}
\mathcal{P}(Q, \nu)(x) &= (\nu P_{Q,\mu,\mu^\alpha})(x) - \nu(x), \\
\mathcal{T}(Q, \mu, \mu^\alpha)(x, a) &= f(x, a, \mu, \mu^\alpha) + \gamma \sum x' p(x'|x, a, \mu, \mu^\alpha) \min_{a'} Q(x', a') - Q(x, a), \\
P_{Q,\mu,\mu^\alpha}(x, x') &= p(x'|x, \arg \min_a Q(x, a), \mu, \mu^\alpha), \quad (\nu P_{Q,\mu,\mu^\alpha})(x) = \sum x_0 \nu(x_0) P_{Q,\mu,\mu^\alpha}(x_0, x).
\end{align*}
\]
ODEs from the three scales $\rho_k^\mu \ll \rho_k^Q \ll \rho_k^{\mu \alpha}$

We choose $\rho_k^\mu \ll \rho_k^Q$ so that $\rho_k^\mu / \rho_k^Q$ is of order $\epsilon \ll 1$, and $\rho_k^Q \ll \rho_k^{\mu \alpha}$ so that $\rho_k^Q / \rho_k^{\mu \alpha}$ is of order $\tilde{\epsilon} \ll 1$.

The following ODE system tracks the previous system:

\[
\begin{align*}
\dot{\mu}_t &= \mathcal{P}(Q_t, \mu_t) \\
\dot{Q}_t &= \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t, \mu_{\alpha}^t) \\
\dot{\mu}_{\alpha}^t &= \frac{1}{\epsilon \tilde{\epsilon}} \mathcal{P}(Q_t, \mu_{\alpha}^t)
\end{align*}
\]

For a fixed action table $Q$, we assume that the third ODE has an asymptotically ($\tilde{\epsilon} \to 0$) stable equilibrium $\mu_{\alpha}^Q$ such that $\mathcal{P}(Q, \mu_{\alpha}^Q) = 0$.

The second ODE becomes

\[
\dot{Q}_t = \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t, \mu_{\alpha}^t).
\]

Again, we assume that it has a stable equilibrium ($\epsilon \to 0$), which we call $(Q_\infty, \mu_\infty)$ and which satisfies that $\mathcal{T}(Q_\infty, \mu_\infty, \mu_{\alpha}^{Q_\infty}) = 0$.

From the asymptotic ($t \to \infty$) equilibrium in the first ODE we also have $\mathcal{P}(Q_\infty, \mu_\infty) = 0$, so that $\mu_\infty = \mu_{\alpha}^{Q_\infty}$, which in turns implies that $\mu_\infty$ and the action given by minimizing $Q_{\mu_{\alpha}^\infty}$ solves our mixed Mean Field Control Game.
ODEs from the three scales $\rho^\mu_k \ll \rho^Q_k \ll \rho^\mu_k$

We choose $\rho^\mu_k \ll \rho^Q_k$ so that $\rho^\mu_k / \rho^Q_k$ is of order $\epsilon \ll 1$, and $\rho^Q_k \ll \rho^\mu_k$ so that $\rho^Q_k / \rho^\mu_k$ is of order $\tilde{\epsilon} \ll 1$.

The following ODE system tracks the previous system:

$$
\begin{align*}
\dot{\mu}_t &= \mathcal{P}(Q_t, \mu_t) \\
\dot{Q}_t &= \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t, \mu^\alpha_t) \\
\dot{\mu}^\alpha_t &= \frac{1}{\epsilon \tilde{\epsilon}} \mathcal{P}(Q_t, \mu^\alpha_t)
\end{align*}
$$

For a fixed action table $Q$, we assume that the third ODE has an asymptotically ($\tilde{\epsilon} \rightarrow 0$) stable equilibrium $\mu^\alpha_Q$ such that $\mathcal{P}(Q, \mu^\alpha_Q) = 0$.

The second ODE becomes

$$
\dot{Q}_t = \frac{1}{\epsilon} \mathcal{T}(Q_t, \mu_t, \mu^\alpha_{Q_t}).
$$

Again, we assume that it has a stable equilibrium ($\epsilon \rightarrow 0$), which we call $(Q_\infty, \mu_\infty)$ and which satisfies that $\mathcal{T}(Q_\infty, \mu_\infty, \mu^\alpha_{Q_\infty}) = 0$.

From the asymptotic $(t \rightarrow \infty)$ equilibrium in the first ODE we also have $\mathcal{P}(Q_\infty, \mu_\infty) = 0$, so that $\mu_\infty = \mu^\alpha_{Q_\infty}$, which in turns implies that $\mu_\infty$ and the action given by minimizing $Q_{\mu^\alpha_\infty}$ solves our mixed Mean Field Control Game.
A Benchmark Asymptotic MFCG

We consider a continuous-time and space benchmark linear-quadratic MFCG problem with a running cost given by

\[ f(x, \alpha, \mu, \mu^\alpha) = \frac{1}{2} \alpha^2 + c_1(x - m)^2 + c_3(x - c_4)^2 + \tilde{c}_1(x - m^\alpha)^2 + \tilde{c}_5(m^\alpha)^2, \]

where \( m = \int x d\mu(x) \) and \( m^\alpha = \int x d\mu^\alpha(x) \), \( c_1, \tilde{c}_1 \), and \( \tilde{c}_5 \) are some positive constant. Here, \( \mu \) and \( \mu^\alpha \) are understood as global environment and local environment. The constant \( c_1 \) is global effect and \( \tilde{c}_1, \tilde{c}_5 \) are local effects. The asymptotic formulation of this mixed MFCG problem is given by

\[
\inf_{\alpha} \mathbb{E} \left[ \int_0^\infty e^{-\beta t} \left( \frac{1}{2} \alpha_t^2 + c_1(X_t^\alpha - m)^2 + c_3(X_t^\alpha - c_4)^2 + \tilde{c}_1(X_t^\alpha - m^\alpha)^2 + \tilde{c}_5(m^\alpha)^2 \right) dt \right].
\]

subject to \( dX_t^\alpha = \alpha_t dt + \sigma dW_t, X_0^\alpha \sim \mu_0 \), and the fixed point condition
\( m = \lim_{t \to \infty} \mathbb{E}(X_t^{\hat{\alpha}}) = m^{\hat{\alpha}}, \) where \( \hat{\alpha} \) is the optimal action.
We consider a continuous-time and space benchmark linear-quadratic MFCG problem with a running cost given by

\[ f(x, \alpha, \mu, \mu^{\alpha}) = \frac{1}{2} \alpha^2 + c_1 (x - m)^2 + c_3 (x - c_4)^2 + \tilde{c}_1 (x - m^{\alpha})^2 + \tilde{c}_5 (m^{\alpha})^2, \]

where \( m = \int x d\mu(x) \) and \( m^{\alpha} = \int x d\mu^{\alpha}(x) \), \( c_1, \tilde{c}_1, \) and \( \tilde{c}_5 \) are some positive constant. Here, \( \mu \) and \( \mu^{\alpha} \) are understood as global environment and local environment. The constant \( c_1 \) is global effect and \( \tilde{c}_1, \tilde{c}_5 \) are local effects. The asymptotic formulation of this mixed MFCG problem is given by

\[
\inf_{\alpha} \mathbb{E} \left[ \int_0^\infty e^{-\beta t} \left( \frac{1}{2} \alpha_t^2 + c_1 (X_t^{\alpha} - m)^2 + c_3 (X_t^{\alpha} - c_4)^2 + \tilde{c}_1 (X_t^{\alpha} - m^{\alpha})^2 + \tilde{c}_5 (m^{\alpha})^2 \right) dt \right].
\]

subject to \( dX_t^{\alpha} = \alpha_t dt + \sigma dW_t, X_0^{\alpha} \sim \mu_0 \), and the fixed point condition \( m = \lim_{t \to \infty} \mathbb{E}(X_t^{\hat{\alpha}}) = m^{\hat{\alpha}} \), where \( \hat{\alpha} \) is the optimal action.
Let $V(x) = \inf_{\alpha} J^x(\alpha)$ be the optimal value function conditioned on $X_0 = x$. It satisfies the HJB equation:

$$\beta V(x) - H(x, \alpha, \mu, \mu^\alpha) - \int_\mathbb{R} \frac{\partial H}{\partial \mu^\alpha} H(h, \alpha, \mu, \mu^\alpha)(x) d\mu^\alpha(h) = 0,$$

with the Hamiltonian

$$H(x, \alpha, \mu, \mu^\alpha) = \inf_{\alpha} \{ A^X V(x) + f(x, \alpha, \mu, \mu^\alpha) \}$$

$$= \inf_{\alpha} \left\{ \alpha \dot{V}(x) + \frac{1}{2} \sigma^2 \ddot{V}(x) + \frac{1}{2} \alpha^2 + c_1 (x - m)^2 + c_3 (x - c_4)^2 + \tilde{c}_1 (x - m^\alpha)^2 + \tilde{c}_5 (m^\alpha)^2 \right\}$$

$$= -\frac{1}{2} \dot{V}(x)^2 + \frac{1}{2} \sigma^2 \ddot{V}(x) + c_1 (x - m)^2 + c_3 (x - c_4)^2 + \tilde{c}_1 (x - m^\alpha)^2 + \tilde{c}_5 (m^\alpha)^2,$$

Using the ansatz $V(x) = \Gamma_2 x^2 + \Gamma_1 x + \Gamma_0$, the optimal control is given by

$$\hat{\alpha}(x) = -\dot{V}(x) = -2\Gamma_2 x - \Gamma_1,$$

and, using the fixed point condition $m = m^\alpha$, we derive the explicit formulas

$$\hat{\alpha}(x) = -2\Gamma_2 \left( x - \frac{c_3 c_4}{c_3 + \tilde{c}_5} \right), \quad \Gamma_2 = \frac{-\beta + \sqrt{\beta^2 + 8 (c_1 + c_3 + \tilde{c}_1)}}{4}.$$

Note that $\mu^\alpha \sim N\left( \frac{c_3 c_4}{c_3 + \tilde{c}_5}, \frac{\sigma^2}{4\Gamma_2^2} \right)$ as the limiting distribution of the OU process $(X_t^\hat{\alpha})$. 

Jean-Pierre Fouque (UCSB)
Let $V(x) = \inf_\alpha J^x(\alpha)$ be the optimal value function conditioned on $X_0 = x$. It satisfies the HJB equation:

$$\beta V(x) - H(x, \alpha, \mu, \mu^\alpha) - \int_{\mathbb{R}} \frac{\partial H}{\partial \mu^\alpha} H(h, \alpha, \mu, \mu^\alpha)(x) d\mu^\alpha(h) = 0,$$

with the Hamiltonian

$$H(x, \alpha, \mu, \mu^\alpha) = \inf_\alpha \left\{ A^X V(x) + f(x, \alpha, \mu, \mu^\alpha) \right\}$$

$$= \inf_\alpha \left\{ \alpha \dot{V}(x) + \frac{1}{2} \sigma^2 \ddot{V}(x) + \frac{1}{2} \alpha^2 + c_1 (x - m)^2 + c_3 (x - c_4)^2 + \tilde{c}_1 (x - m^\alpha)^2 + \tilde{c}_5 (m^\alpha)^2 \right\}$$

$$= -\frac{1}{2} \dot{V}(x)^2 + \frac{1}{2} \sigma^2 \ddot{V}(x) + c_1 (x - m)^2 + c_3 (x - c_4)^2 + \tilde{c}_1 (x - m^\alpha)^2 + \tilde{c}_5 (m^\alpha)^2 ,$$

Using the ansatz $V(x) = \Gamma_2 x^2 + \Gamma_1 x + \Gamma_0$, the optimal control is given by

$$\hat{\alpha}(x) = -\dot{V}(x) = -2\Gamma_2 x - \Gamma_1 ,$$

and, using the fixed point condition $m = m^\alpha$, we derive the explicit formulas

$$\hat{\alpha}(x) = -2\Gamma_2 \left( x - \frac{c_3 c_4}{c_3 + \tilde{c}_5} \right) , \quad \Gamma_2 = \frac{-\beta + \sqrt{\beta^2 + 8 (c_1 + c_3 + \tilde{c}_1)}}{4} .$$

Note that $\mu^{\hat{\alpha}} = \mathcal{N} \left( \frac{c_3 c_4}{c_3 + \tilde{c}_5}, \frac{\sigma^2}{4\Gamma_2^2} \right)$ as the limiting distribution of the OU process $(X_t^{\hat{\alpha}})$. 
Results from the RL algorithm after discretization in \((x, a)\)

Parameters: \(c_1 = 0.25\), \(c_3 = 0.5\), \(c_4 = 0.6\) \(\tilde{c}_1 = 0.3\), \(\tilde{c}_5 = 5\), discount rate \(\beta = 1\), and volatility of the state dynamics \(\sigma = 0.3\).

With \(\rho = (1 + k)^{-\omega}\), the learning rates are \((\omega^{\mu}, \omega^{Q}, \omega^{\mu^{\alpha}}) = (0.85, 0.55, 0.15)\), and \(\epsilon = 0.05\) for the \(\epsilon\)-greedy policy. We run the algorithm with \(K = 80000\) episodes.

The blue line is the theoretical control function and the blue dots are the learned control. The green curve shows the theoretical distribution of state, where the global distribution equals to the local distribution. The orange (resp. violet) curve refers to the empirical global (resp. local) distribution learned by the algorithm.
Trader’s problem revisited

Suppose we have $M$ companies, each company with $N$ traders following the same strategy. The stock inventory $X_{t}^{m,n}$ of the $m$-th company’s $n$-th trader follows the dynamics,

$$dX_{t}^{m,n} = \alpha_{t}^{m}(X_{t}^{m,n})dt + \sigma_{t}^{m,n}dW_{t}^{m,n},$$

and her cash $K_{t}^{m,n}$ evolves as

$$dK_{t}^{m,n} = -[\alpha_{t}^{m}(X_{t}^{m,n})S_{t} + c_{\alpha}(\alpha_{t}^{m}(X_{t}^{m,n}))]dt,$$

where $c_{\alpha}(\cdot)$ is a positive function for trading cost.

The stock price is impacted with a function $h(\cdot)$ by all transactions

$$dS_{t} = \frac{1}{M} \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N} h(\alpha_{t}^{m}(X_{t}^{m,n}))dt + \sigma_{t}^{0}dW_{t}^{0}.$$ 

The total wealth of the individual trader $(m, n)$ is

$$V_{t}^{m,n} = K_{t}^{m,n} + X_{t}^{m,n}S_{t}$$
Trader’s problem revisited

Suppose we have $M$ companies, each company with $N$ traders following the same strategy. The stock inventory $X_{t}^{m,n}$ of the $m$-th company’s $n$-th trader follows the dynamics,

$$dX_{t}^{m,n} = \alpha_{t}^{m}(X_{t}^{m,n})dt + \sigma_{t}^{m,n}dW_{t}^{m,n},$$

and her cash $K_{t}^{m,n}$ evolves as

$$dK_{t}^{m,n} = - [\alpha_{t}^{m}(X_{t}^{m,n})S_{t} + c_{\alpha}(\alpha_{t}^{m}(X_{t}^{m,n}))] dt,$$

where $c_{\alpha}(\cdot)$ is a positive function for trading cost.

The stock price is impacted with a function $h(\cdot)$ by all transactions

$$dS_{t} = \frac{1}{M} \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N} h(\alpha_{t}^{m}(X_{t}^{m,n}))dt + \sigma_{t}^{0}dW_{t}^{0}.$$

The total wealth of the individual trader $(m, n)$ is

$$V_{t}^{m,n} = K_{t}^{m,n} + X_{t}^{m,n}S_{t}$$
Trader’s problem revisited

The dynamics of the individual trader’s wealth is:

\[
dV_t^{m,n} = dK_t^{m,n} + S_t dX_t^{m,n} + X_t^{m,n} dS_t
\]

\[=
\left[-c_\alpha \left(\alpha_t^m (X_t^{m,n})\right) + X_t^{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} h \left(\alpha_t^i (X_t^{i,j})\right)\right] dt + S_t \sigma_t^{m,n} dW_t^{m,n} + X_t^{m,n} \sigma_0^0 dW_0^0.
\]

We assume that the individual trader is subject to a running liquidation constraint modeled by a function \(c_X\) of the shares equally held by her own group. In this model, the individual trader’s objective function is given by

\[
J^{m,n} (\alpha_1^1, \ldots, \alpha_M) = \mathbb{E} \left\{ \int_0^T c_X \left( \frac{1}{N} \sum_{j=1}^{N} X_t^{m,j} \right) dt + g \left( X_T^{m,n} \right) - V_T^{m,n} \right\}
\]

\[=
\mathbb{E} \left\{ \int_0^T \left[ c_X \left( \frac{1}{N} \sum_{j=1}^{N} X_t^{m,j} \right) + c_\alpha \left(\alpha_t^m (X_t^{m,n})\right) - X_t^{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} h \left(\alpha_t^i (X_t^{i,j})\right)\right] dt + g \left( X_T^{m,n} \right) \right\}.
\]

In the limit of a large number of large groups (without precising the relation between \(M\) and \(N\)), and assuming \(\sigma_t^{m,n} = \sigma\), this problem leads to the following mixture of MFG and MFC problems: Minimize

\[
J (\alpha) = \mathbb{E} \left\{ \int_0^T \left[ c_X \left( m_t^\alpha \right) + c_\alpha \left(\alpha_t^m\right) - X_t \int h(a) d\theta_t (a)\right] dt + g \left( X_T \right) \right\},
\]

where \(\theta_t\) is the law of the control \(\alpha_t\), \(m_t^\alpha = \mathbb{E} (X_t^\alpha)\), and

\[
dX_t = \alpha_t dt + \sigma dW_t, \quad t \leq T, \quad X_0 = x.
\]

Note that the problem is a MFG of control through \(\theta_t\) and a MFC of the state through \(m_t^\alpha\).
Trader’s problem revisited

The dynamics of the individual trader’s wealth is:

\[
\begin{align*}
\frac{dV_{t}^{m,n}}{dt} &= \frac{dK_{t}^{m,n}}{dt} + S_{t}dX_{t}^{m,n} + X_{t}^{m,n}dS_{t} \\
&= \left[-c_{\alpha} \left(\alpha_{t}^{m} (X_{t}^{m,n})\right) + X_{t}^{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} h \left(\alpha_{t}^{i} (X_{t}^{i,j})\right)\right] dt + S_{t}\sigma_{t}^{m,n} dW_{t}^{m,n} + X_{t}^{m,n} \sigma_{0}^{0} dW_{0}^{0}.
\end{align*}
\]

We assume that the individual trader is subject to a running liquidation constraint modeled by a function \( c_X \) of the shares equally held by her own group. In this model, the individual trader’s objective function is given by

\[
J_{t}^{m,n} (\alpha^{1}, \ldots, \alpha^{M}) = \mathbb{E} \left\{ \int_{0}^{T} c_{X} \left(\frac{1}{N} \sum_{j=1}^{N} X_{t}^{m,j}\right) dt + g \left( X_{T}^{m,n} \right) - V^{m,n}_{T} \right\}
\]

\[
= \mathbb{E} \left\{ \int_{0}^{T} \left[ c_{X} \left(\frac{1}{N} \sum_{j=1}^{N} X_{t}^{m,j}\right) + c_{\alpha} \left(\alpha_{t}^{m} (X_{t}^{m,n})\right) - X_{t}^{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} h \left(\alpha_{t}^{i} (X_{t}^{i,j})\right)\right] dt + g \left( X_{T}^{m,n} \right) \right\}.
\]

In the limit of a large number of large groups (without precising the relation between \( M \) and \( N \)), and assuming \( \sigma_{t}^{m,n} = \sigma \), this problem leads to the following mixture of MFG and MFC problems: Minimize

\[
J (\alpha) = \mathbb{E} \left\{ \int_{0}^{T} \left[ c_{X} \left(m_{t}^{\alpha}\right) + c_{\alpha} \left(\alpha_{t}\right) - X_{t} \int h(\alpha) d\theta_{t}(\alpha)\right] dt + g \left( X_{T}\right) \right\},
\]

where \( \theta_{t} \) is the law of the control \( \alpha_{t} \), \( m_{t}^{\alpha} = \mathbb{E} (X_{t}^{\alpha}) \), and

\[
dX_{t} = \alpha_{t} dt + \sigma dW_{t}, \quad t \leq T, \quad X_{0} = x.
\]

Note that the problem is a MFG of control through \( \theta_{t} \) and a MFC of the state through \( m_{t}^{\alpha} \).
Trader’s problem revisited

The dynamics of the individual trader’s wealth is:

\[
\frac{dV_{m,n}}{dt} = K_{m,n} + X_{m,n} S t dX_{m,n} + X_{m,n} dS_t
\]

\[
= \left[ -c_\alpha (\alpha_t (X_{m,n})) + X_{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^M \sum_{j=1}^N h (\alpha_t (X_{i,j})) \right] dt + S_t \sigma_{m,n} dW_{m,n} + X_{m,n} \sigma_0 dW_0.
\]

We assume that the individual trader is subject to a running liquidation constraint modeled by a function \( c_X \) of the shares equally held by her own group. In this model, the individual trader’s objective function is given by

\[
J_{m,n} (\alpha) = E \left\{ \int_0^T c_X \left( \frac{1}{N} \sum_{j=1}^N X_{m,j} \right) dt + g(X_T) - V_{m,n} \right\}
\]

\[
= E \left\{ \int_0^T \left[ c_X \left( \frac{1}{N} \sum_{j=1}^N X_{m,j} \right) + c_\alpha (\alpha_t (X_{m,n})) - X_{m,n} \frac{1}{M} \frac{1}{N} \sum_{i=1}^M \sum_{j=1}^N h (\alpha_t (X_{i,j})) \right] dt + g(X_T) \right\}.
\]

In the limit of a large number of large groups (without precising the relation between \( M \) and \( N \)), and assuming \( \sigma_{m,n} = \sigma \), this problem leads to the following mixture of MFG and MFC problems: Minimize

\[
J (\alpha) = E \left\{ \int_0^T \left[ c_X (m_\alpha_t) + c_\alpha (\alpha_t) - X_t \int h(a) d\theta_t(a) \right] dt + g(X_T) \right\},
\]

where \( \theta_t \) is the law of the control \( \alpha_t \), \( m_\alpha_t = \mathbb{E}(X_{m,n}^{\alpha}) \), and

\[
dX_t = \alpha_t dt + \sigma dW_t, \quad t \leq T, \quad X_0 = x.
\]

Note that the problem is a MFG of control through \( \theta_t \) and a MFC of the state through \( m_\alpha_t \).
Trader’s problem revisited

In what follows we specialize to the **Linear-Quadratic** case where $c_x(m) = \frac{c_X}{2} m^2$, $c_{\alpha}(\alpha) = \frac{c_{\alpha}}{2} \alpha^2$, $h(a) = c_h a$, and $g(x) = \frac{c_g}{2} x^2$, so that:

$$J(\alpha) = \mathbb{E} \left\{ \int_0^T \left[ \frac{c_X}{2} m^2_t + \frac{c_{\alpha}}{2} \alpha^2_t - c_h X_t \int a d\theta_t(a) \right] dt + \frac{c_g}{2} X_T^2 \right\}.$$  

In order to solve this problem, one first “freezes” the flow $(\theta_t)$ (MFG), and then solves the control problem which is of MKV type due to the term $m_t = \mathbb{E}(X_t)$ (MFC). Differentiating the corresponding Hamiltonian with respect to $\alpha$, one gets

$$\hat{\alpha}_t = - \frac{1}{c_{\alpha}} Y_t.$$

On the other hand,

$$dY_t = - \left( -c_h \mathbb{E}[\hat{\alpha}_t] + c_X \mathbb{E}[X_t] \right) dt + Z_t dW_t,$$

which leads to the following FBSDE:

$$\begin{cases}
    dX_t = - \frac{1}{c_{\alpha}} Y_t dt + \sigma dW_t, & X_0 \sim \mu_0, \\
    dY_t = - \left( \frac{c_h}{c_{\alpha}} \mathbb{E}[Y_t] + c_X \mathbb{E}[X_t] \right) dt + Z_t dW_t, & Y_T = c_g X_T.
\end{cases}$$
Trader’s problem revisited

In what follows we specialize to the **Linear-Quadratic** case where \( c_x(m) = \frac{c_x}{2} m^2 \), \( c_\alpha(\alpha) = \frac{c_\alpha}{2} \alpha^2 \), \( h(a) = c_h a \), and \( g(x) = \frac{c_g}{2} x^2 \), so that:

\[
J(\alpha) = \mathbb{E} \left\{ \int_0^T \left[ \frac{c_x}{2} m_t^2 + \frac{c_\alpha}{2} \alpha_t^2 - c_h X_t \int ad\theta_t(a) \right] dt + \frac{c_g}{2} X_T^2 \right\}.
\]

In order to solve this problem, one first “freezes” the flow \((\theta_t)\) (MFG), and then solves the control problem which is of MKV type due to the term \( m_t = \mathbb{E}(X_t) \) (MFC). Differentiating the corresponding Hamiltonian with respect to \( \alpha \), one gets

\[
\hat{\alpha}_t = -\frac{1}{c_\alpha} Y_t.
\]

On the other hand,

\[
dY_t = -(-c_h \mathbb{E}[\hat{\alpha}_t] + c_x \mathbb{E}[X_t]) dt + Z_t dW_t,
\]

which leads to the following FBSDE:

\[
\begin{cases}
  dX_t = -\frac{1}{c_\alpha} Y_t dt + \sigma dW_t, \quad X_0 \sim \mu_0, \\
  dY_t = - \left( \frac{c_h}{c_\alpha} \mathbb{E} [Y_t] + c_x \mathbb{E} [X_t] \right) dt + Z_t dW_t, \quad Y_T = c_g X_T.
\end{cases}
\]
Trader’s problem revisited: analytic solution

Taking expectation in this system one obtains:

\[
\begin{align*}
    d\mathbb{E}[X_t] &= -\frac{1}{c^\alpha} \mathbb{E}[Y_t] dt, \quad \mathbb{E}[X_0] = x_0, \\
    d\mathbb{E}[Y_t] &= -\left(\frac{\gamma}{c^\alpha} \mathbb{E}[Y_t] + cX \mathbb{E}[X_t]\right) dt, \quad \mathbb{E}[Y_T] = c_g \mathbb{E}[X_T].
\end{align*}
\]

Solving this system leads to

\[
\mathbb{E}[Y_t] = \bar{\eta}(t) \mathbb{E}[X_t],
\]

where

\[
\bar{\eta}_t = \frac{-C(e^{(\delta^+ - \delta^-)}(T-t) - 1) - c_g(\delta^+ e^{(\delta^+ - \delta^-)}(T-t) - \delta^-)}{(\delta^- e^{(\delta^+ - \delta^-)}(T-t) - \delta^+) - c_g B(e^{(\delta^+ - \delta^-)}(T-t) - 1)},
\]

for \( t \in [0, T] \), \( B = 1/c^\alpha \), \( C = cX \), \( \delta^\pm = -D \pm \sqrt{R} \), with \( D = -c_h/(2c^\alpha) \) and \( R = D^2 + BC \).

Subsequently:

\[
\mathbb{E}[X_t] = x_0 e^{-\int_0^t \frac{\bar{\eta}(s)}{c^\alpha} ds}.
\]

From the FBSDE system for \((X_t, Y_t, Z_t)\) and centering \(X_t\) and \(Y_t\), one gets:

\[
Y_t = \eta(t)X_t + \psi(t), \quad \eta(t) = \frac{c\alpha c_g}{c\alpha + c_g(T-t)}, \quad Z_t = \sigma \eta(t), \quad \psi(t) = (\bar{\eta}(t) - \eta(t)) \mathbb{E}[X_t].
\]

Finally, we recall that the optimal control is given by:

\[
\hat{\alpha}_t = -\frac{1}{c^\alpha} Y_t = -\frac{1}{c^\alpha} (\eta(t)X_t + \psi(t)).
\]

Assuming that \(X_0\) is \( \mathcal{N}(x_0, \sigma_0^2) \)-distributed and independent of \(W\), \(X_t\) is normally-distributed with mean given above by \(\mathbb{E}[X_t] = x_0 e^{-\int_0^t \frac{\bar{\eta}(s)}{c^\alpha} ds}\) and variance \(Var(X_t) = \sigma_0^2 e^{-\frac{2}{c^\alpha} \int_0^t \eta(s) ds} + \sigma^2 \int_0^t e^{-\frac{2}{c^\alpha} \int_s^t \eta(s') ds'} ds\) easily computed from

\[
dX_t = -\frac{1}{c^\alpha} (\eta(t)X_t + \psi(t)) dt + \sigma dW_t.
\]
Numerical Results for the Trader’s Problem

- Parameters: $c_\alpha = 1$, $c_X = 0.75$, $c_h = 1.25$, $c_g = 1$, and with a volatility for the state dynamic $\sigma = 0.75$. The distribution of the initial inventory $X_{t_0}$ is $\mathcal{N}(0.25, 0.5)$.

- The terminal time is $T = 1$, and we choose a time grid $\tau = \{t_0 = 0, \ldots, t_{NT} = T\}$ with time step $\Delta t = 1/16$.

- We discretize the state space into $\mathcal{X} = \{x_0 = -2, \ldots, x_{|\mathcal{X}|-1} = 2.5\}$, and the action space into $\mathcal{A} = \{a_0 = -2, \ldots, a_{|\mathcal{A}|-1} = 1.5\}$, where the step sizes are $\Delta x = \Delta a = \sqrt{\Delta t} = 1/4$.

- The triplet of the learning rates is chosen as $(\omega^\theta, \omega^Q, \omega^\mu) = (0.85, 0.55, 0.15)$. For the $\epsilon$-greedy policy we choose $\epsilon = 0.05$.

- We run the experiment 10 times, each with $K = 200000$ episodes. We average the control and state distributions learned by Algorithm over the last 10000 episodes and over 10 runs.
We have introduced a type of mixed **Mean Field Control Games (MFCG)** that model competitive games between a large number of large collaborative groups.

It turns out that the two timescales Reinforcement Learning algorithm U2-MF-QL introduced in AFL2021 for infinite horizon problems and in AFL2022 for finite horizon extended problems, for learning either MFG or MFC problems, is naturally adapted for learning MFCG problems by managing three learning rates in the U3-MF-QL algorithm proposed here (ADFL2022).

We illustrate the results with linear quadratic problems for which we have explicit formulas. In particular, a new type of trader’s problem is presented.

The theory for MFCGs is a work in progress, as well as an actor-critic version of the U3-MF-QL algorithm in the context of continuous spaces.
Conclusion

- We have introduced a type of mixed **Mean Field Control Games (MFCG)** that model competitive games between a large number of large collaborative groups.

- It turns out that the two timescales Reinforcement Learning algorithm U2-MF-QL introduced in AFL2021 for infinite horizon problems and in AFL2022 for finite horizon extended problems, for learning either MFG or MFC problems, is naturally adapted for learning MFCG problems by managing three learning rates in the **U3-MF-QL algorithm** proposed here (ADFL2022).

- We illustrate the results with linear quadratic problems for which we have explicit formulas. In particular, a new type of trader’s problem is presented.

- The theory for MFCGs is a **work in progress**, as well as an actor-critic version of the U3-MF-QL algorithm in the context of continuous spaces.
We have introduced a type of mixed Mean Field Control Games (MFCG) that model competitive games between a large number of large collaborative groups.

It turns out that the two timescales Reinforcement Learning algorithm U2-MF-QL introduced in AFL2021 for infinite horizon problems and in AFL2022 for finite horizon extended problems, for learning either MFG or MFC problems, is naturally adapted for learning MFCG problems by managing three learning rates in the U3-MF-QL algorithm proposed here (ADFL2022).

We illustrate the results with linear quadratic problems for which we have explicit formulas. In particular, a new type of trader’s problem is presented.

The theory for MFCGs is a work in progress, as well as an actor-critic version of the U3-MF-QL algorithm in the context of continuous spaces.
We have introduced a type of mixed Mean Field Control Games (MFCG) that model competitive games between a large number of large collaborative groups.

It turns out that the two timescales Reinforcement Learning algorithm U2-MF-QL introduced in AFL2021 for infinite horizon problems and in AFL2022 for finite horizon extended problems, for learning either MFG or MFC problems, is naturally adapted for learning MFCG problems by managing three learning rates in the U3-MF-QL algorithm proposed here (ADFL2022).

We illustrate the results with linear quadratic problems for which we have explicit formulas. In particular, a new type of trader’s problem is presented.

The theory for MFCGs is a work in progress, as well as an actor-critic version of the U3-MF-QL algorithm in the context of continuous spaces.
Thank you all for your attention and stay healthy