Learning Mean-Field Games

Junzi Zhang
Stanford ICME
e-mail: junziz@stanford.edu

Joint work with Prof. Xin Guo, Anran Hu and Renyuan Xu
UC Berkeley, IEOR
Outline

Mathematical Framework
- Motivating Problem
- General $N$-player game and GMFG
- RL for $N = 1$

GMFG with RL
- Existence and Uniqueness of GMFG solution
- Convergence and Complexity of RL
- Numerical Performance
Outline

Mathematical Framework
  Motivating Problem
  General $N$-player game and GMFG
  RL for $N = 1$

GMFG with RL
  Existence and Uniqueness of GMFG solution
  Convergence and Complexity of RL
  Numerical Performance
Mathematical Framework

Motivating Problem
General $N$-player game and GMFG
RL for $N = 1$

GMFG with RL
Existence and Uniqueness of GMFG solution
Convergence and Complexity of RL
Numerical Performance
Motivation: a sequential auction game

Ad auction problem for advertisers:

- **Ad auction**: a stochastic game on an ad exchange platform among a large number of players (the advertisers)
- **Environment**: in each round, a web user requests a page, and then a Vickrey-type *second-best-price* auction is run to incentivize advertisers to bid for a slot to display advertisement
- **Characteristics**:
  - partial information (unknown conversion of clicks, unknown bid price of other competitors)
  - changing states: budget constraint

**Question**: how should one bid in this sequential game with a large population of competing bidders and unknown distributions of the conversion of clicks/rewards and bids/actions of other bidders?
Motivation: sequential auction game

Solution: the simultaneous learning and decision-making problem in a sequential auction with a large number of homogeneous bidders.

- Full model approach: solve it as an $N$-player game
  - multi-agent reinforcement learning: computationally intractable

- Approximation approaches:
  - independent learners (regarding others as environment) (IL)
  - multi-agent reinforcement learning with first-order (expectation) mean-field approximation (MF-Q, Yang et al., 2018)

- Our approach: Reinforcement Learning (RL) + full distribution Mean-Field Game (MFG) approximation
Outline

**Mathematical Framework**
- Motivating Problem
- General $N$-player game and GMFG
- RL for $N = 1$

**GMFG with RL**
- Existence and Uniqueness of GMFG solution
- Convergence and Complexity of RL
- Numerical Performance
Overview of MFG

**Mean-Field Game (MFG)** is
- a game with very large population of small interacting individuals
  - **large population**: a continuum of players
  - **small interacting**: strategy based on the aggregated macroscopic information (mean field)
- originated from physics on weakly interacting particles
- theoretical works pioneered by Lasry and Lions (2007) and Huang, Malhamé and Caines (2006)
Main Idea of MFG

- Take an $N$-player game;
- When $N$ is large, consider instead the “aggregated” version of the $N$-player game;
- By (f)SLLN, the aggregated version, MFG, becomes an “approximation” of the $N$-player game, in terms of $\epsilon$-Nash equilibrium.
N-player game

\[
\begin{align*}
\text{maximize}_{\pi_i} & \quad V^i(s, \pi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r^i(s_t, a^i_t) \mid s^0 = s \right] \\
\text{subject to} & \quad s^i_{t+1} \sim P^i(s_t, a^i_t)
\end{align*}
\]

- \( N \) players, state space \( S \), action space \( A \);
- \( s_t = (s^1_t, \ldots, s^N_t) \in S^N \) is the state vector;
- \( a_t = (a^1_t, \ldots, a^N_t) \in A^N \) is the action vector;
- admissible (Markovian) policy \( \pi_i : S^N \rightarrow \mathcal{P}(A) \), with \( \mathcal{P}(X) \) the space of all probability measures over \( X \);
- \( r^i \) is the reward function for player \( i \);
- \( P^i \) is the transition dynamics for player \( i \);
- \( \gamma \) is the discount factor;
$N$-player Games

**Definition ($N$-player game: Nash equilibrium (NE))**

NE is a set of strategies such that no agent can benefit from unilaterally deviating from this set of strategies. Formally, $\pi^*$ is an NE if for all $i$ and $s$,

$$V^i(s, \pi^*) \geq V^i(s, (\pi^*_1, \ldots, \pi_i, \ldots, \pi^*_N))$$

holds for any $\pi_i : S^N \rightarrow \mathcal{P}(A)$. 

### From $N$-player Game to MFG

#### $N$-player game

| maximize $V^i(s, \pi)$ := $\mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r^i(s_t, a^i_t) | s_0 = s \right]$ |
| subject to $s_{t+1}^i \sim P^i(s_t, a^i_t)$ |

Assume identical, indistinguishable and interchangeable players. When the number of players goes to infinity, view the limit of $s_{-i}^t = (s_1^t, \ldots, s_{i-1}^t, s_{i+1}^t, \ldots, s_N^t)$ as population state distribution $\mu_t$.

#### MFG

| maximize $V(s, \pi, \{\mu_t\}_{t=0}^{\infty}) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mu_t) | s_0 = s \right]$ |
| subject to $s_{t+1} \sim P(s_t, a_t, \mu_t)$ |

Mean-Field Games (MFG)

MFG

maximize\(\pi\) \(V(s, \pi, \{\mu_t\}_{t=0}^{\infty}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mu_t) | s_0 = s\right]\)

subject to \(s_{t+1} \sim P(s_t, a_t, \mu_t)\).

- infinite number of homogeneous players, state space \(S\), action space \(A\);
- \(s_t \in S\) and \(a_t \in A\) are the state and action of a representative agent at time \(t\);
- \(\mu_t \in \mathcal{P}(S)\) is the population state distribution at time \(t\);
- admissible policy \(\pi : S \times \mathcal{P}(S) \to \mathcal{P}(A)\);
- \(r\) is the reward function, \(P\) is the transition dynamics.
Mean-Field Games (MFG)

**Definition (Stationary NE for MFGs)**

In MFGs, a pair \((\pi^*, \mu^*)\) is called a stationary NE if

1. *(Single agent side)* For any policy \(\pi\) and any initial state \(s \in S\), we have
   \[
   V(s, \pi^*, \{\mu^*\}_{t=0}^\infty) \geq V(s, \pi, \{\mu^*\}_{t=0}^\infty).
   \]

2. *(Population side)* \(\mathbb{P}_{s_t} = \mu^*\) for all \(t \geq 0\), where \(\{s_t\}_{t=0}^\infty\) is the dynamics under control \(\pi^*\) starting from \(s_0 \sim \mu^*\), with \(a_t \sim \pi^*(s_t, \mu^*)\), \(s_{t+1} \sim P(\cdot|s_t, a_t, \mu^*)\).
General $N$-player Games

$N$-player game

maximize $V^i(s, \pi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r^i(s_t, a^i_t) | s_0 = s \right]$

subject to $s^i_{t+1} \sim P^i(s_t, a^i_t)$.

General $N$-player game

maximize $V^i(s, \pi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r^i(s_t, a^i_t) | s_0 = s \right]$

subject to $s^i_{t+1} \sim P^i(s_t, a^i_t)$

$\triangleright a_t = (a^1_t, \cdots, a^N_t)$. 
### General $N$-player Games

#### $N$-player game

| maximize $V_i(s, \pi) := \mathbb{E}\left[ \sum_{t=0}^{\infty} \gamma^t r_i(s_t, a^i_t) \mid s_0 = s \right]$ |
| subject to $s^i_{t+1} \sim P^i(s_t, a^i_t)$ |

#### General $N$-player game

| maximize $V_i(s, \pi) := \mathbb{E}\left[ \sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s \right]$ |
| subject to $s^i_{t+1} \sim P^i(s_t, a^i_t)$ |

$\triangleright\ a_t = (a^1_t, \ldots, a^N_t)$. 

Generalized Mean-Field Games (GMFG)

MFG

\[
\begin{align*}
\text{maximize}_\pi & \quad V(s, \pi, \{\mu_t\}_{t=0}^{\infty}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, \mu_t) \mid s_0 = s\right] \\
\text{subject to} & \quad s_{t+1} \sim P(s_t, a_t, \mu_t).
\end{align*}
\]

GMFG

\[
\begin{align*}
\text{maximize}_\pi & \quad V(s, \pi, \{L_t\}_{t=0}^{\infty}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, L_t) \mid s_0 = s\right] \\
\text{subject to} & \quad s_{t+1} \sim P(s_t, a_t, L_t).
\end{align*}
\]

- \(L_t \in \Delta^{|S||A|}\) is the population state-action pair distribution at time \(t\), with state marginal \(\mu_t\) and action marginal \(\alpha_t\) (population action distribution);
- \(\alpha_t\) as an approximation of \(a_t^{-i} = (a_1^t, \ldots, a_{i-1}^t, a_{i+1}^t, \ldots, a_N^t)\).
Nash Equilibrium in GMFGs

Definition (Stationary NE for GMFGs)

In GMFGs, an agent-population pair \((\pi^*, L^*)\) is called a stationary NE if

1. (Single agent side) For any policy \(\pi\) and any initial state \(s \in S\), we have
   \[
   V(s, \pi^*, \{L^*\}_{t=0}^\infty) \geq V(s, \pi, \{L^*\}_{t=0}^\infty).
   \]

2. (Population side) \(P_{s_t, a_t} = L^*\) for all \(t \geq 0\), where \(\{s_t, a_t\}_{t=0}^\infty\) is the dynamics under control \(\pi^*\) starting from \(s_0 \sim \mu^*\), with \(a_t \sim \pi^*(s_t, \mu^*), s_{t+1} \sim P(\cdot|s_t, a_t, L^*),\) and \(\mu^*\) being the population state marginal of \(L^*\).
Outline

Mathematical Framework
- Motivating Problem
- General $N$-player game and GMFG
- RL for $N = 1$

GMFG with RL
- Existence and Uniqueness of GMFG solution
- Convergence and Complexity of RL
- Numerical Performance
Reinforcement learning: Overview

▶ Single agent problem with unknown $P$ and $r$

$$\text{maximize}_{\pi} \quad V(s, \pi) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right],$$

subject to

$$s_{t+1} \sim P(s_t, a_t), \quad a_t \sim \pi(s_t), \quad t \geq 0.$$  

▶ Simultaneous decision making of $a_t$ and learning of $r$ and $P$, optimal value $V^*(s) := \max_{\pi} V(s, \pi)$

▶ Examples: Chess/Go/Poker
Existing Algorithms for RL

- **Discrete state and action spaces:**
  - Q-learning (Mnih, Kavukcuoglu, Silver, Graves, Antonoglou, Wierstra, & Riedmiller, 2013)
  - PSRL (Osband, Russo & Van Roy, 2013)
  - UCRL2 (Jaksch, Ortner & Auer, 2010)

- **Continuous state and action spaces:**
  - Policy gradient (Williams, 1992)
  - Actor-Critic (Konda & Tsitsiklis, 2000)
  - Linear Quadratic Regulator (LQR): Abbasi-Yadkori & Szepesvári, 2011; Dean, Mania, Matni, Recht, Tu, 2018
Q-learning

- \textbf{Q-function: } $Q^*(s, a) := \mathbb{E}r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')$

- \textbf{Bellman equation (for } Q\text{-function):}

$$Q^*(s, a) = \mathbb{E}r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} \max_{a'} Q^*(s', a')$$

- \textbf{Q-learning: } stochastic approximation to the Bellman equation:

$$Q^{k+1}(s, a) \leftarrow (1 - \beta_t(s, a)) Q^k(s, a) + \beta_t(s, a) \left[ r(s, a) + \gamma \max_{a'} Q^k(s', a') \right]$$
Q-learning

- **Q-function:** \( Q^*(s, a) := \mathbb{E}r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s') \)

- **Bellman equation (for Q-function):**
  \[
  Q^*(s, a) = \mathbb{E}r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} \max_{a'} Q^*(s', a')
  \]

- **Q-learning: stochastic approximation to the Bellman equation:**
  \[
  Q^{k+1}(s, a) \leftarrow (1 - \beta_t(s, a)) Q^k(s, a) + \beta_t(s, a) \left[ r(s, a) + \gamma \max_{a'} Q^k(s', a') \right]
  \]
Key gradients in Q-learning

- With finite state and action spaces, $Q^k$ are matrices
- Choice of appropriate $\beta_t(s, a)$ and exploration in $a$:
  - $\epsilon$-greedy: $a_k \in \arg \max Q^k(s_k, a)$ with probability $1 - \epsilon$, and $a_k$ chosen randomly from $A$ with probability $\epsilon$
  - Boltzmann policy: based on a softmax operator parameterized by $c$
- $Q^k \rightarrow Q^*$
Outline

Mathematical Framework
- Motivating Problem
- General $N$-player game and GMFG
- RL for $N = 1$

GMFG with RL
- Existence and Uniqueness of GMFG solution
- Convergence and Complexity of RL
- Numerical Performance
(Recall) Nash Equilibrium in GMFGs

Definition (Stationary NE for GMFGs)

In GMFGs, an agent-population pair \((\pi^*, L^*)\) is called a stationary NE if

1. **(Single agent side)** For any policy \(\pi\) and any initial state \(s \in S\), we have

\[
V (s, \pi^*, \{L^*\}_{t=0}^\infty) \geq V (s, \pi, \{L^*\}_{t=0}^\infty).
\]

2. **(Population side)** \(\mathbb{P}_{s_0, a_0} = L^*\) for all \(t \geq 0\), where \(\{s_t, a_t\}_{t=0}^\infty\) is the dynamics under control \(\pi^*\) starting from \(s_0 \sim \mu^*\), with \(a_t \sim \pi^*(s_t, \mu^*)\), \(s_{t+1} \sim P(\cdot|s_t, a_t, L^*)\), and \(\mu^*\) being the population state marginal of \(L^*\).
fixed point/three-step approach

- Step 1: given $L$, solve the stochastic control problem to get $\pi^*_L$:

$$
\begin{align*}
\text{maximize}_{\pi} \quad & V(s, \pi, L) := \mathbb{E}\left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, L) | s_0 = s \right], \\
\text{subject to} \quad & s_{t+1} \sim P(s_t, a_t, L).
\end{align*}
$$

- Step 2: given $\pi^*_L$, update from $L$ for one time step to get $L'$ following the dynamics.

- Step 3: Check whether $L'$ matches $L$, and repeat.
Mappings $\Gamma_1$ and $\Gamma_2$

- Take any fixed population action-state distribution $L \in \mathcal{P}(S \times A)$,
  \[
  \Gamma_1 : \mathcal{P}(S \times A) \rightarrow \Pi := \{\pi \mid \pi : S \rightarrow \mathcal{P}(A)\},
  \]
  such that $\pi^*_L = \Gamma_1(L)$ is an optimal policy given $L$.

- For any admissible policy $\pi \in \Pi$ and $L \in \mathcal{P}(S \times A)$, define $\Gamma_2 : \Pi \times \mathcal{P}(S \times A) \rightarrow \mathcal{P}(S \times A)$ as
  \[
  \Gamma_2(\pi, L) := L' = \mathbb{P}_{s_1, a_1},
  \]
  where $a_1 \sim \pi(s_1)$, $s_1 \sim \mu P(\cdot|\cdot, a_0, L)$, $a_0 \sim \pi(s_0)$, $s_0 \sim \mu$, and $\mu$ is the population state marginal of $L$. 
Existence and Uniqueness

Theorem 1 (Guo, Hu, Xu, & Zhang, 2019)

For any GMFG, if $\Gamma_2 \circ \Gamma_1$ is contractive, then there exists a unique stationary NE. In addition, the three-step approach converges.

Remark 1: Here the uniqueness is in the sense of $L$.
Remark 2: Similar assumption and result can be found in (Huang, Malhamé & Caines, 2006) for MFGs.
Remark 3: We indeed established Theorem 1 in much more general settings without directly assuming contractivity, and we allow for

- non-stationarity, general compact state and action spaces, and Wasserstein metrics.

See our draft for more details.

Question: How to solve the GMFG when there is uncertainty in $r$ and $P$? Assume in the following that $\mathcal{S}$ and $\mathcal{A}$ are both finite.
Existence and Uniqueness

Theorem 1 (Guo, Hu, Xu, & Zhang, 2019)

For any GMFG, if $\Gamma_2 \circ \Gamma_1$ is contractive, then there exists a unique stationary NE. In addition, the three-step approach converges.

Remark 1: Here the uniqueness is in the sense of $L$.

Remark 2: Similar assumption and result can be found in (Huang, Malhamé & Caines, 2006) for MFGs.

Remark 3: We indeed established Theorem 1 in much more general settings without directly assuming contractivity, and we allow for

- non-stationarity, general compact state and action spaces, and Wasserstein metrics.

See our draft for more details.

Question: How to solve the GMFG when there is uncertainty in $r$ and $P$? Assume in the following that $S$ and $A$ are both finite.
Three-step approach revisited:

- **Step 1:** given $L$, solve the stochastic control problem to get $\pi^*_L$:
  
  \[
  \text{maximize}_{\pi} \quad V(s, \pi, L) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, L) | s_0 = s \right], \\
  \text{subject to} \quad s_{t+1} \sim P(s_t, a_t, L).
  \]

- **Step 2:** given $\pi^*_L$, update from $L$ for one time step to get $L'$ following the dynamics.

- **Step 3:** Check whether $L'$ matches $L$. 

Three-step approach revisited (when $P$ and $R$ are unknown):

- **Step 1:** given $L$, solve a RL problem with transition dynamics $P_L(s'|s, a) := P(s'|s, a, L)$ and reward $r_L(s, a) := r(s, a, L)$ via Q-learning:

  \[
  Q^{k+1}_L(s, a) \leftarrow (1 - \beta_t(s, a))Q^k_L(s, a) + \beta_t(s, a) \left[ r(s, a, L) + \gamma \max_{a'} Q^k_L(s', a') \right].
  \]

- **Step 2:** given $\pi^*_L$, update from $L$ for one time step to get $L'$ following the dynamics.

- **Step 3:** Check whether $L'$ matches $L$.

**Remark:** $\pi^*_L(s) \in \arg\max_a Q^*_L(s, a)$. When $\arg\max$ is non-unique, replace it with $\arg\max-e$, which assigns equal probability to the maximizers.
Algorithm 1 Naive Q-learning for GMFGs

1: \textbf{Input}: Initial population state-action pair $L_0$
2: \textbf{for} $k = 0, 1, \cdots$ \textbf{do}
3: \hspace{1em} Perform Q-learning to find the Q-function $Q_k^*(s, a) = Q_{L_k}^*(s, a)$ of an MDP with dynamics $P_{L_k}(s'|s, a)$ and reward distributions $R_{L_k}(s, a)$.
4: \hspace{1em} Solve $\pi_k \in \Pi$ with $\pi_k(s) = \text{argmax}_e (Q_k^*(s, \cdot))$.
5: \hspace{1em} Sample $s \sim \mu_k$, where $\mu_k$ is the population state marginal of $L_k$, and obtain $L_{k+1}$ from $G(s, \pi_k, L_k)$.
6: \textbf{end for}
Failure of the Naive Algorithm

Failure examples:

(a) fluctuation in $l_\infty$.

(b) fluctuation in $l_1$.

Figure: Fluctuations of Naive Algorithm (30 sample paths).
Algorithm 1 Naive Q-learning for GMFGs

1: **Input**: Initial population state-action pair $L_0$
2: **for** $k = 0, 1, \cdots$ **do**
3: Perform Q-learning to find the Q-function $Q_k^*(s, a) = Q_{L_k}^*(s, a)$ of an MDP with dynamics $P_{L_k}(s'|s, a)$ and reward distributions $R_{L_k}(s, a)$. 
4: Solve $\pi_k \in \Pi$ with $\pi_k(s) = \text{argmax-e} (Q_k^*(s, \cdot))$.
5: Sample $s \sim \mu_k$, where $\mu_k$ is the population state marginal of $L_k$, and obtain $L_{k+1}$ from $G(s, \pi_k, L_k)$. 
6: **end for**
Instability of \texttt{argmax-e}:

Magnify the Approximation Errors

\begin{itemize}
\item $x = (1, 1)$, then $\texttt{argmax-e}(x) = (1/2, 1/2)$.
\item $y = (1, 1 - \epsilon)$, then for any $\epsilon > 0$, $\texttt{argmax-e}(y) = (1, 0)$.
\item $\|\texttt{argmax-e}(x) - \texttt{argmax-e}(y)\|_2 / \|x - y\|_2 = 1/\epsilon$ – non-Lipschitz.
\end{itemize}
Instability of argmax-e:

Magnify the Approximation Errors

- $x = (1, 1)$, then $\text{argmax-e}(x) = (1/2, 1/2)$.
- $y = (1, 1 - \epsilon)$, then for any $\epsilon > 0$, $\text{argmax-e}(y) = (1, 0)$.
- $\|\text{argmax-e}(x) - \text{argmax-e}(y)\|_2 / \|x - y\|_2 = 1/\epsilon$ - non-Lipschitz.
Instability of $\text{argmax-e}$:

Magnify the Approximation Errors

- $x = (1, 1)$, then $\text{argmax-e}(x) = (1/2, 1/2)$.
- $y = (1, 1 - \epsilon)$, then for any $\epsilon > 0$, $\text{argmax-e}(y) = (1, 0)$.
- $\|\text{argmax-e}(x) - \text{argmax-e}(y)\|_2 / \|x - y\|_2 = 1/\epsilon$ – non-Lipschitz.
Algorithm 2 Q-learning for GMFGs (GMF-Q)

1: **Input**: Initial $L_0$, tolerance $\epsilon > 0$.
2: **for** $k = 0, 1, \cdots$ **do**
3: Perform Q-learning for $T_k$ iterations to find the approximate Q-function $\hat{Q}_k^* (s, a) = \hat{Q}^*_{L_k} (s, a)$ of an MDP with dynamics $P_{L_k} (s' | s, a)$ and reward distributions $R_{L_k} (s, a)$.
4: Compute $\pi_k \in \Pi$ with $\pi_k (s) = \text{softmax}_c (\hat{Q}_k^* (s, \cdot))$.
5: Sample $s \sim \mu_k$, where $\mu_k$ is the population state marginal of $L_k$, and obtain $\tilde{L}_{k+1}$ from $G(s, \pi_k, L_k)$.
6: Find $L_{k+1} = \text{Proj}_{S_\epsilon} (\tilde{L}_{k+1})$
7: **end for**

**Remark.** Here $S_\epsilon$ is a $\epsilon$-net of $L$, and $\text{softmax}_c (x)_i = \frac{\exp(c x_i)}{\sum_{j=1}^{n} \exp(c x_j)}$. 
Outline

Mathematical Framework
  Motivating Problem
  General $N$-player game and GMFG
  RL for $N = 1$

GMFG with RL
  Existence and Uniqueness of GMFG solution
  Convergence and Complexity of RL
  Numerical Performance
Theorem 2 (Guo, Hu, Xu, & Zhang, 2019)

Given the same assumptions in the existence and uniqueness theorem, for any specified tolerances $\epsilon, \delta > 0$, with appropriate choices of $T_k, c$ and $S_\epsilon$, \( \limsup_{k \to \infty} W_1(L_k, L^*) = O(\epsilon) \) with probability at least $1 - 2\delta$.

Here $W_1$ is the $\ell_1$ Wasserstein distance, a.k.a. earth mover distance.
Complexity of MF-AQ

Theorem 3 (Guo, Hu, Xu. & Zhang, 2019)

Given the same assumptions in the existence and uniqueness theorem, for any specified tolerances $\epsilon$, $\delta > 0$, set $T_k$, $c$ and $S_\epsilon$ appropriately. Then with probability at least $1 - 2\delta$, $W_1(L_{K_\epsilon}, L^*) = O(\epsilon)$, and the total number of iterations $T = \sum_{k=0}^{K_\epsilon-1} T_k$ is bounded by

$$T = O \left( K_\epsilon^{19/3} (\log(K_\epsilon/\delta))^{41/3} \right).$$

Here $K_\epsilon := \left\lceil 2 \max \left\{ (\eta\epsilon)^{-1/\eta}, \log_d(\epsilon / \max\{\text{diam}(S)\text{diam}(A), 1\}) + 1 \right\} \right\rceil$ is the number of outer iterations.
Outline

Mathematical Framework
  Motivating Problem
  General $N$-player game and GMFG
  RL for $N = 1$

GMFG with RL
  Existence and Uniqueness of GMFG solution
  Convergence and Complexity of RL
  Numerical Performance
Repeated Auction Example Revisited

At each round $t$:

- randomly select $M - 1$ players (from $N$, possibly infinite players) to compete with the representative advertiser
- $a_{t}^{M}$: second best price among the bids from $M$ players
- reward $r_{t} = \mathbf{1}_{w_{t}^{M}=1} \left[ (v_{t} - a_{t}^{M}) - (1 + \rho) \mathbf{1}_{s_{t}<a_{t}^{M}}(a_{t}^{M} - s_{t}) \right]$
  
  - $v_{t}$: conversion
  - $w_{t}$: indicator of winning (bid the highest price)
  - $s_{t}$: current budget
  - $\rho$: penalty of overbidding
- dynamic of the budget:
  
  $$s_{t+1} = \begin{cases} 
  s_{t}, & w_{t} \neq 1, \\
  s_{t} - a_{t}^{M}, & w_{t} = 1 \text{ and } a_{t}^{M} \leq s_{t}, \\
  0, & w_{t} = 1 \text{ and } a_{t}^{M} > s_{t}.
  \end{cases}$$

- Budget fulfillment: modify the dynamics of $s_{t+1}$ with a non-negative random budget fulfillment $\Delta(s_{t+1})$ after the auction clearing, such that $\hat{s}_{t+1} = s_{t+1} + \Delta(s_{t+1})$. 
Performance against full-information

When transition $P$ and reward $r$ are known, replace Q-learning with value iteration (VI) – GMF-V.

$$Q_{L}^{k+1}(s,a) \leftarrow \mathbb{E}r(s,a,L) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q_{L}^{k}(s',a'),$$

<table>
<thead>
<tr>
<th>$T_{k}^{\text{GMF-Q}}$</th>
<th>1000</th>
<th>3000</th>
<th>5000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta Q$</td>
<td>0.21263</td>
<td>0.1294</td>
<td>0.10258</td>
<td>0.0989</td>
</tr>
</tbody>
</table>

Table: Q-table with $T_{k}^{\text{GMF-V}} = 5000$.

Here $\Delta Q := \frac{\|Q_{\text{GMF-V}} - Q_{\text{GMF-Q}}\|_{2}^{2}}{\|Q_{\text{GMF-V}}\|_{2}^{2}}$ is the relative $L_{2}$ distance between the Q-tables.
Performance against full-information

(a) GMF-Q.  
(b) GMF-V.

Figure: Q-tables: GMF-Q vs. GMF-V. 20 outer iterations.

Conclusion: our algorithm (requiring no specific information on $P$ and $R$) can learn almost as well as algorithms with full information.
Performance against S.O.T.A.

Performance metric:

\[
C(\pi) = \frac{1}{N|S|^N} \sum_{i=1}^{N} \sum_{s \in S^N} \frac{\max_{\pi_i} V_i(s, (\pi^{-i}, \pi^i)) - V_i(s, \pi)}{|\max_{\pi_i} V_i(s, (\pi^{-i}, \pi^i))| + \epsilon_0}.
\]

Here \(\epsilon_0 > 0\) is a safeguard, and is taken as 0.1 in the experiments. If \(\pi^*\) is an NE, by definition, \(C(\pi^*) = 0\) and it is easy to check that \(C(\pi) \geq 0\).
Performance against S.O.T.A.

Compare our GMF-Q with IL (independent learners) and MF-Q ($N$-player game with first-order mean-field approximation, Yang et al., 2018).

**Figure:** Learning accuracy based on $C(\pi)$. $|\mathcal{S}| = |\mathcal{A}| = 10$, $N = 20$. 90% confidence interval, 20 sample paths.
Performance against S.O.T.A.

Compare our GMF-Q with IL (independent learners) and MF-Q ($N$-player game with first-order mean-field approximation, Yang et al., 2018).

Figure: Learning accuracy based on $C(\pi)$. $|S| = |A| = 20, N = 20$. 90% confidence interval, 20 sample paths.
Performance against S.O.T.A.

Compare our GMF-Q with IL (independent learners) and MF-Q (\(N\)-player game with first-order mean-field approximation, Yang et al., 2018).

**Figure:** Learning accuracy based on \(C(\pi)\). \(|S| = |A| = 10, N = 40\). 90\% confidence interval, 20 sample paths.
Conclusions

In this work, we

- build a generalized mean-field games framework with learning in a MFG;
- establish the unique existence for the GMFG solution for the discrete time version;
- propose a Q-learning algorithm with convergence and complexity analysis;
- numerical experiments demonstrate superior performance compared to existing RL algorithms.
Thank you!

Reference:
