MF-OMO: An Optimization Formulation of Mean-Field Games

Xin Guo
UC Berkeley and Amazon
xinguo@berkeley.edu

Joint work with Anran Hu (U. of Oxford) and Junzi Zhang (Citedal)

Stochastic Control and Financial Engineering

Princeton University, June 20, 2023
Outline

Motivating Examples and Challenges

MF-OMO: MFGs as Occupation Measure Optimization

Solving MF-OMO: Convergence to NE solutions

Numerics
Outline

Motivating Examples and Challenges

MF-OMO: MFGs as Occupation Measure Optimization

Solving MF-OMO: Convergence to NE solutions

Numerics
Motivating example: a sequential auction game

Example: Bid recommendation in sequential ad auction

search query $\Longrightarrow$ (hybrid) auction

Characteristics:

- A large number of almost identical advertisers (over 1MM)
- Multiple stakeholders of possibly conflicting interests: Amazon/Google/Meta vs advertisers, vs consumers
- Coexistence of competition and collaboration

\[^1\text{Figure source: https://mwmstudioz.com/facebook-ad-auction/}\]
Motivating example: rideshare

- A large number of drivers
- Multiple stakeholders of possibly conflicting interests: Uber, Lyft vs drivers, vs passengers
- Coexistence of competition and collaboration

Figure source: https://360.here.com/2014/04/30/jams-game-theory-equations-science-of-traffic/
Multi-level multi-agent problems

Many problems, including online advertising and healthcare
- multi-agent games with large number of agents
- multiple stakeholders involved with possibly conflicting goals
- coexistence of competition and collaborations
First technical challenge: large population

Resolution: scale down the interaction to the weak form

- mean-field approximation
  ⇒ a single representative agent’s interaction with the empirical distribution of a large homogeneous population
Main challenge: characterizing NEs

- existence of multiple NEs
- interplay between PO and NEs: optimization problem for the central planner with the constraint of NEs

Remark: different parties may have conflicting interests, and followers may not follow the leader, unlike the principle-agent problems
Extended/Generalized MFG

- Considers a representative agent with state $s_t$, action $a_t$, and solves

$$\max_{\pi} \mathbb{E} \left[ \sum_{t=0}^{T} r_t(s_t, a_t, L_t) \big| s_0 \sim \mu_0 \right]$$

subject to

$$s_{t+1} \sim P_t(s_t, a_t, L_t), \quad a_t \sim \pi_t(\cdot | s_t), \quad t = 0, \ldots, T - 1$$

- Finite-time-horizon $\mathcal{T} = \{0, 1, \ldots, T\}$
- Finite state and action spaces $S$ and $A$
- Population state-action joint distribution $L_t \in \Delta(S \times A)$
- Reward and transition probability $r_t, P_t$
- Mixed policy/Relaxed control: $\pi_t : S \to \Delta(A)$
Nash equilibrium of MFG

A pair \((\pi, L)\) is called a Nash equilibrium (NE) for the MFG if it satisfies

### Optimal \(\pi\) given \(L\) (best response/optimality))

\[
\text{maximize}_{\pi} \quad \mathbb{E} \left[ \sum_{t=0}^{T} r_t(s_t, a_t, L_t) \mid s_0 \sim \mu_0 \right] \\
\text{subject to} \quad s_{t+1} \sim P_t(s_t, a_t, L_t), \quad a_t \sim \pi_t(\cdot | s_t), \quad t = 0, \ldots, T - 1
\]

### Consistent \(L\) given \(\pi\) (Fokker-Planck (consistency))

\[
L_0(s, a) = \mu_0(s)\pi_0(a | s), \\
L_{t+1}(s', a') = \pi_{t+1}(a' | s') \sum_{s \in S} \sum_{a \in A} L_t(s, a) P_t(s' | s, a, L_t), \quad \forall t = 0, \ldots, T - 1
\]
Existence of Nash equilibrium of MFG

Proposition

Suppose that $P_t(s'|s, a, L_t)$ and $r_t(s, a, L_t)$ are both continuous in $L_t$ for any $s, s' \in S$, $a \in A$ and $t \in T$. Then a Nash equilibrium solution exists. Moreover, this Nash equilibrium solution is an $\epsilon$-Nash equilibrium to the original $N$-player game.

How to find some or all Nash equilibria?
Solution approaches for continuous-time MFGs

- PDE/control approach: backward HJB equation + forward Kolmogorov equation
  *Lions and Lasry (2007), Huang, Malhame and Caines (2006), Lions, Lasry and Guánt (2009),*

- Probabilistic approach: FBSDEs
  *Buckdahn, Li and Peng (2009), Carmona and Delarue (2013)*

- Master equation (and verification argument)
  *Cardaliaguet, Delarue, Lasry, and Lions (2019)*
Basic idea of existing approaches:

Given $L = \{L_t\}_{t=0}^T$, finding the optimal control/best response is an MDP (or RL), denoted as $\mathcal{M}(L)$.

- **Alternate** between representative agent and population mean-field
  - Step 1: given $L$ (and optionally $\pi$), use $\mathcal{M}(L)$ to get an updated control/policy $\pi'$
  - Step 2: given $\pi'$, update from $L$ for one time step to get $L'$ following the dynamics
  - Step 3: Check whether $L'$ matches $L$, and repeat
Some existing computation algorithms

- **GMF-Q (Guo, H., Xu & Zhang (2019))**
  - $Q^{k+1} = Q^*_L$
  - $\pi^{k+1} = \text{softmax}_{c_{k+1}}(Q^{k+1})$
  - $L^{k+1} = \text{MeanFieldUpdate}(\pi^{k+1})$

- **Ficticious Play (Perrin et al. (2020))**
  - $\pi^{k+1} \in \text{BestResponse}(\tilde{L}^k)$
  - $L^{k+1} = \text{MeanFieldUpdate}(\pi^{k+1})$
  - $\tilde{L}^{k+1} = \alpha_{k+1}\tilde{L}^k + (1 - \alpha_{k+1})L^{k+1}$ (e.g., $\alpha_{k+1} = \frac{k}{k+1}$)

- **Online Mirror Descent (Perolat et al. (2021))**
  - $Q^{k+1} = Q^\pi_L$, $\tilde{Q}^{k+1} = \tilde{Q}^k + \alpha Q^{k+1}$
  - $\pi^{k+1} = \text{softmax}_c(\tilde{Q}^k)$
  - $L^{k+1} = \text{MeanFieldUpdate}(\pi^{k+1})$
Existing assumptions

- **Contractivity**: Mapping from $L$ to $L'$ in Best Response algorithm is contractive

  - Guo, H., Xu & Zhang (2019, 2020), Anahtarci, Kariksiz & Saldi (2019, 2020), Fu, Yang, Chen & Wang (2019), Wang, Han, Yang & Wang (2020), Xie, Yang, Wang & Minca (2020), Cui & Koppel (2021), Subramanian, Taylor, Crowley & Poupart (2022)

- **Monotonicity**: For any $L, L' \in \Delta(S \times A)$,

  $$\sum_{s \in S} \sum_{a \in A} (L(s, a) - L'(s, a))(r(s, a, L) - r(s, a, L')) \leq 0.$$

  - Elie et al. (2019), Perrin et al. (2020), Lee, Rengarajan, Kalathil & Shakkottai (2020), Perolat et al. (2021), Geist et al. (2021), Perrin et al. (2021)

- **Uniqueness of Nash equilibrium**
Remarks

- Assumptions needed to find Nash equilibrium are stronger than the assumptions for the existence of Nash equilibrium
- Contractive or monotonicity conditions generally do not hold for games
- Most games have more than one Nash equilibrium

**Our focus:** finding multiple NEs of MFGs?
Outline

Motivating Examples and Challenges

MF-OMO: MFGs as Occupation Measure Optimization

Solving MF-OMO: Convergence to NE solutions

Numerics
Focus on an MDP $\mathcal{M}$ (e.g., $\mathcal{M}(L)$)

- **Occupation measures**: a sequence $\{d_t\}_{t \in T} \subseteq \Delta(S \times A)$
  - $d_t(s, a) = \mathbb{P}^{\mu_0, \pi}(s_t = s, a_t = a)$

- Define the set-valued mapping $\Pi$ which maps from a sequence $\{d_t\}_{t \in T}$ to a set of policy/control sequences $\{\pi_t\}_{t \in T}$:

$$\pi_t(a|s) = \frac{d_t(s, a)}{\sum_{a' \in A} d_t(s, a')} \quad (*)$$

when $\sum_{a' \in A} d_t(s, a') > 0$, and $\pi_t(\cdot|s)$ is an arbitrary probability vector in $\Delta(A)$ when $\sum_{a' \in A} d_t(s, a') = 0$
MDP as LP of occupation measures

**Lemma (Guo, H. & Zhang (2022))**

\( \{\pi_t\}_{t \in T} \) is an \( \epsilon \)-suboptimal policy/control of the MDP \( \mathcal{M} \) if and only if \( d_t(s, a) = \mu_t(s)\pi_t(a|s) \) is a feasible \( \epsilon \)-suboptimal solution to the following linear program:

Maximize
\[
\sum_{t \in T} \sum_{s \in S} \sum_{a \in A} d_t(s, a)r_t(s, a)
\]
Subject to
\[
\sum_{s \in S} \sum_{a \in A} d_t(s, a)P_t(s'|s, a) = \sum_{a \in A} d_{t+1}(s', a), \quad \forall s' \in S, t \in T \setminus \{T\},
\]
\[
\sum_{a \in A} d_0(s, a) = \mu_0(s), \quad \forall s \in S,
\]
\[
d_t(s, a) \geq 0, \quad \forall s \in S, a \in A, t \in T.
\]
LP formulation of MDPs/stochastic controls/MFGs

Discrete-time models:
- MDPs with finite state-action spaces: Manne (1960)
- Approximate dynamic programming: Schweitzer and Seidmann (1985), De Farias and Van Roy (2003)

Continuous-time modes:
- Semi-MDPs: Osaki and Mine (1968)
- Controlled martingale problems: Stockbridge (1990), Bhatt and Borkar (1996), Kurtz and Stockbridge (1998)
- Singular controls: Taksar (1997), Kurtz and Stockbridge (2001)
- Optimal stopping in MFGs: Bouveret, Dumitrescu and Tankov (2020), Dumitrescu, Leutscher and Tankov (2021, 2022)
Second component: duality and consistency

\textbf{(A)} Optimal control/Best response: \( d = \{d_t(s, a)\}_{t \in \mathcal{T}} \) solves the following linear program

\[
\begin{align*}
&\text{maximize}_{x} \quad \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} x_t(s, a) r_t(s, a, L_t) \\
&\text{subject to} \quad \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} x_t(s, a) P_t(s' | s, a, L_t) = \sum_{a \in \mathcal{A}} x_{t+1}(s', a), \quad \forall s' \in \mathcal{S}, t \in \mathcal{T} \setminus \{T\}, \\
&\quad \sum_{a \in \mathcal{A}} x_0(s, a) = \mu_0(s), \quad \forall s \in \mathcal{S}, \\
&\quad x_t(s, a) \geq 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}
\end{align*}
\]

\textbf{(B)} Consistency of population flow:

\[
\begin{align*}
L_0(s, a) &= \mu_0(s) \pi_0(a | s), \\
L_{t+1}(s', a') &= \pi_{t+1}(a' | s') \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} L_t(s, a) P_t(s' | s, a, L_t), \quad \forall t = 0, \ldots, T - 1,
\end{align*}
\]

where \( \pi \in \Pi(d) \)
Key observation 1: The LP in Condition A can be written as the following

\[
\begin{align*}
\text{minimize}_{d} & \quad c_{L}^\top d \\
\text{subject to} & \quad A_{L}d = b, \quad d \geq 0.
\end{align*}
\]

Here matrix $A_{L}$ is defined by transition probabilities $P$, vector $c_{L}$ is defined by the reward functions $r$ and vector $b$ is defined by the initial distribution $\mu_0$. 
More precisely, \( b = [0, \ldots, 0, \mu_0] \in \mathbb{R}^{ST+S} \),

\[
c_L = [-r_0(\cdot, \cdot, L_0), \ldots, -r_T(\cdot, \cdot, L_T)] \in \mathbb{R}^{SA(T+1)},
\]

\[
A_L = \begin{bmatrix}
W_0(L_0) & -Z & 0 & 0 & \cdots & 0 & 0 \\
0 & W_1(L_1) & -Z & 0 & \cdots & 0 & 0 \\
0 & 0 & W_2(L_2) & -Z & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & W_{T-1}(L_{T-1}) & -Z \\
Z & 0 & 0 & 0 & \cdots & 0 & 0
\end{bmatrix}.
\]

Here \( W_t(L_t) \in \mathbb{R}^{S \times SA} \) is the matrix with the \( l \)-th row \( (l = 1, \ldots, S) \) being the flattened vector \( [P_t(l| \cdot, \cdot, L_t)] \in \mathbb{R}^{SA} \), and \( Z := [I_S, \ldots, I_S] \in \mathbb{R}^{S \times SA} \).
Strong duality

The following conditions are equivalent:

- **Optimal control/Best response condition:** $d^*$ solves

  $\minimize_d \ c_L^T d$

  subject to \quad $A_L d = b, \quad d \geq 0$

- **There exist** $y, z$ such that

  $A_L d^* = b, \quad A_L^T y + z = c_L,$

  $z^T d^* = 0, \quad d^* \geq 0, \quad z \geq 0$
Key observation 2: Given $d = \{d_t(s, a)\}_{t \in \mathcal{T}}$ satisfies condition A, condition B can be reduced to the following condition

$$(B') \quad d_t(s, a) = L_t(s, a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathcal{T}$$

Interpretation: single agent occupation measure = population mean field
Theorem (Guo, H. & Zhang (2022))

Solving a Nash equilibrium of the mean-field game problem is equivalent to solving the following feasibility optimization problem.

\[
\begin{align*}
\text{minimize}_{y,z,L} & \quad 0 \\
\text{subject to} & \quad A_L L = b, \quad A_L^\top y + z = c_L, \\
& \quad z^\top L = 0, \quad L \geq 0, \quad z \geq 0.
\end{align*}
\]

If \((\pi, L)\) is an NE of the mean-field game, then there exist some \(y, z\) such that \((y, z, L)\) is a feasible solution to the above optimization problem; Conversely, if \((y, z, L)\) is a feasible solution to the above optimization problem, then \((\pi, L)\) is an NE of the mean-field game, with \(\forall \pi \in \Pi(L)\) by (*)..

- \(y\) is the value function; \(z\) is the Bellman residual
- In fact, there is indeed no loss to restrict \(y\) and \(z\) to be bounded and interpretable.
Exploitability measure the closeness of a policy to NE:

- $\text{Expl}(\pi)$ quantifies the gain for a representative player to replace its policy/control by a best response/optimal control, while the rest of the population plays with policy $\pi$.
- **Fact**: $\text{Expl}(\pi) \geq 0$ and $\text{Expl}(\pi) = 0$ if and only if $\pi$ is an NE.
- $\text{Expl}(\pi) \leq \varepsilon$ characterizes $\varepsilon$-NE.
Solving (MF-OMO) exactly is equivalent to solving for NE.

To find an $\epsilon$-NE, it is sufficient to approximately solve MF-OMO to $O(\epsilon^2)$ sub-optimality.

\[
\begin{align*}
\text{minimize}_{y,z,L} & \quad \|A_L L - b\|_2^2 + \|A_L^T y + z - c_L\|_2^2 + z^T L \\
\text{subject to} & \quad L \geq 0, \quad 1^T L = 1, \\
& \quad 1^T z \leq SA(T^2 + T + 2)r_{\max}, \quad z \geq 0, \\
& \quad \|y\|_2 \leq \frac{S(T+1)(T+2)}{2}r_{\max}.
\end{align*}
\]
Outline

Motivating Examples and Challenges

MF-OMO: MFGs as Occupation Measure Optimization

Solving MF-OMO: Convergence to NE solutions

Numerics
Solving MF-OMO: Convergence to NE solutions

Optimization variable: \( \theta := (y, z, L) \).

Projected gradient descent:

\[ \theta_{k+1} = \text{Proj}(\theta_k - \eta_k \nabla_{\theta} f_{\text{MF-OMO}}(\theta_k)) \].

Theorem (Guo, H. & Zhang (2022))

Under some regularity assumption 1, projected gradient descent algorithm can solve (MFOMO) with local convergence guarantee.
Solving MF-OMO: convergence to NE solutions

Consider a special class of MFGs:

- $r_t$ is linear in $L_t$: $r_t(s, a, L_t) = \bar{r}_{s,a,t} + b_{t,s,a}^\top \bar{R}_{t,s,a}$
- $P_t$ is independent of $L_t$: $P_t(s'|s, a, L_t) = \bar{P}_{s',s,a,t}$

Finding NE solution(s) of this class of MFGs is equivalent to solving a linear complementarity problem:

$$\begin{align*}
\text{minimize}_{y,z,L} & : \begin{bmatrix} 0 & A_L & y \\ A_L^\top & -B & L \end{bmatrix} + \begin{bmatrix} 0 \\ z \end{bmatrix} = \begin{bmatrix} b \\ \bar{c} \end{bmatrix}, \\
\text{subject to} & : z^\top L = 0, \quad L \geq 0, \quad z \geq 0,
\end{align*}$$

which can be reduced to solving a finite number of linear programs.

Proposition (Guo, H. & Zhang (2022))

*Suppose that the MFG has linear rewards and mean-field independent dynamics. Then its NE solution can be found in finite time.*
Outline

Motivating Examples and Challenges

MF-OMO: MFGs as Occupation Measure Optimization

Solving MF-OMO: Convergence to NE solutions

Numerics
SIS: comparison of algorithms

- Large/infinite number of agents choose to either social distance or to go out at each time step
- Susceptible agents get infected when going out, with probability proportional to the number of infected agents; otherwise agents stay healthy
- Infected agents recover with some probability at each step
- Agents aim at finding the best strategy to minimize their costs induced from social distancing and being infected
SIS example with $T=50$: (normalized) exploitability vs. iteration

Figure: Convergence against iterations
Figure: Convergence against runtime
Convergence results (1/2)

- Consider an MFG with at least three known NEs
- $\delta$ is the initialization neighborhood around each NE
- With random initializations around the $\delta$ neighborhood of NE $i$, we record the proportion of convergence to 1e-2 and 1e-3 normalized exploitability under 400 iterations.

| $\delta$ | NE 1 | NE 2 | NE 3 |
|----------|------|------|------|
|         | 1e-2 | 1e-3 | 1e-2 | 1e-3 | 1e-2 | 1e-3 |
| 0.05     | 100% | 100% | 100% | 95%  | 100% | 100% |
| 0.2      | 100% | 100% | 100% | 95%  | 100% | 95%  |
| 0.8      | 100% | 95%  | 100% | 95%  | 100% | 100% |
| 1.0      | 100% | 100% | 100% | 100% | 100% | 100% |

Table: Convergence behavior of different tolerances
With randomly sampled from the $\delta$ neighborhood of NE $i$, we record the proportion of convergence to some NE under 400 iterations.

$p_i$ is the proportion of convergence closest to NE $i$.

| $\delta$ | NE 1  | NE 2  | NE 3  |
|-----------|-------|-------|-------|
|           | $p_1$ | $p_2 + p_3$ | $p_2$ | $p_1 + p_3$ | $p_3$ | $p_1 + p_2$ |
| 0.05      | 70%   | 30%   | 75%   | 20%   | 65%   | 35%   |
| 0.2       | 50%   | 50%   | 65%   | 30%   | 60%   | 35%   |
| 0.8       | 45%   | 50%   | 40%   | 55%   | 30%   | 70%   |
| 1.0       | 40%   | 60%   | 45%   | 55%   | 25%   | 75%   |

Table: Convergence behavior of different initializations: Tolerance 1e-3
This reformulation of NE sets as feasibility of LP has been generalized to analyze multi-level multi-agent systems, including

- generalization to ergodic, infinite-time horizon and more general forms of MFGs
- stability and sensitivity issues for Stackelberg mean field games (G., Hu, and Zhang (2022))
- balancing social optimality and NE for bid recommendation systems (G., Li, Nabi, Salhab, Zhang (2023))
- continuous time-state spaces (on going)

X. Guo, A. Hu, J. Zhang (2022).
MF-OMO: An Optimization Formulation of Mean-Field Games
arxiv.org/abs/2206.09608
MFGLib: an open-source Python library for MFGs.

- **Environments:**
  - Pre-implemented envs
  - Off-the-shelf interface for user-defined envs

- **Solvers:**
  - MFOMO, OnlineMirrorDescent, PriorDescent, FictitiousPlay
  - Embedded auto-tuning tool to automatically select the best algorithm hyperparameters

- Available at [https://github.com/radar-research-lab/MFGLib](https://github.com/radar-research-lab/MFGLib).

Comments and suggestions are welcome!
Thank you!
Definability

Assumption

For any \( s \in S, a \in A \), \( P_t(s, a, L_t) \) and \( r_t(s, a, L_t) \) (as functions of \( L_t \)) are both restrictions of definable functions on the log-exp structure to \( \Delta(S \times A) \).

The precise definition of definability and the log-exp structure can be found in Attouch, Bolte, Redont & Soubeyran (2010), Section 4.3.

Examples:

- All semi-algebraic functions, all analytic functions on compact sets as well as the exponential and logarithm functions are definable.
- Any finite combination of definable functions via summation, subtraction, product, affine mapping, composition, (generalized) inversion, partial supremum and partial infimum, as well as reciprocal (restricted to a compact domain) inside their domain of definitions are definable.