Welfare Maximization in Competitive Equilibrium: Reinforcement Learning for Markov Exchange Economy

Zhihan Liu[1], Miao Lu[2], Zhaoran Wang[1], Michael Jordan [3], and Zhuoran Yang[4]

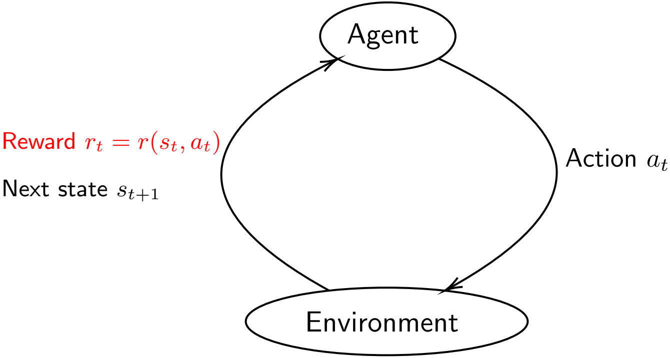
[1]Northwestern University [2]University of Science and Technology of China, [3]UC Berkeley, [4]Yale University

June 7, 2023

Exchange Economy and Social Welfare Maximization

- In *exchange economy (EE)*, a set of rational agents with individual initial endowments allocate and exchange a finite set of valuable resources based on a common price system.
- The target of EE is to achieve Competitive Equilibrium (CE), where all agents maximize their own utilities *under their budget constraint*.
- When each agent within a system is to *myopically* maximize its own utility at each step, a *central planner* is introduced to steer the system so as to achieve *Social Welfare Maximization (SWM)*.

Reinforcement Learning



- The agent aims to learn a policy π which maximizes its state value function $V_1^\pi(s_1)$ at the first step and the initial state s_1.
- State value function $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{h=1}^{H} r(s_h, a_h) \mid s_h = s]$.

Challenges

- Problem formulation and optimality characterization of a *dynamic bilevel economic system* involving both EE and SWM.
- Exploration-exploitation tradeoff in online learning and distribution shift in offline learning.
- Adoption of *general function approximation*.

Main Contribution

- We propose a new economic system known as Markovian Exchange Economy (MEE) and define a suboptimality function for the planner and the agents.
- For online and offline MEE, we design MARL-style algorithms, proving the online regret and the offline suboptimality, respectively.

Markovian Exchange Economy (MEE)

- A finite horizon MEE consists of N agents, one social planner, and H time steps.
- Each state s_h consists a context c_h and endowments e_h.
- The joint actions of the agents consist the allocations for each agent and the price for the exchange.
- **Interaction Protocol:** At each time step $h \in [H]$, the agents and the planner observe state $s_h^k \in \mathcal{S}$ and pick their own actions a_h^k and b_h^k. Then the next state is generated by the environment $s_{h+1}^k \sim P_h(\cdot \mid s_h^k, b_h^k)$ and they observe the utilities $\{u_h^{k,(i)}\}_{i\in[N]}$ with $u_h^{k,(i)} = u_h^{(i)}(s_h^k, x_h^{k,(i)})$ from the environment.

Characterization of Optimality

Agent policy $\nu : \mathcal{S} \mapsto \mathcal{A}, s \mapsto (\nu^{(1)}(s), \cdots, \nu^{(N)}(s), \nu^{\mathbf{p}}(s))$.

- Optimality: one-step *competitive equilibrium* (Definition 2.2).
- Characterized by a fixed-point formulation for value functions (Theorem 2.4).

Planner policy $\pi : \mathcal{S} \mapsto \mathcal{B}, s \mapsto \pi(s)$.

- Optimality: *maximize social welfare* (sum of utilities).
- Characterized by another fixed-point formulation for value functions (Theorem 2.6).

Joint optimality: policy pair $(\pi^\star, \nu^\star)$ satisfying *competitive equilibrium* and *social welfare maximization* simultaneously.

- Planner's policy π is coupled with agents' policy ν.
- Fixed-point formulation (Theorem 2.7) $\Rightarrow$ Suboptimality of any policy pair (π, ν), denoted by $\mathrm{SubOpt}(\nu, \pi)$.

Model-based Optimistic online Learning for MEE (MOLM)

MOLM algorithm design (two steps):

- **Model estimation step:** construct confidence sets $\mathcal{U}_h^k$ for utility functions and $\mathcal{P}_h^k$ for transition kernels using data from previous $k-1$ episodes.
- We use value targeted regression (VTR, Ayoub et al., 2020) for transition estimation.
- **Optimistic planning step:** use $\mathcal{U}_h^k$ and $\mathcal{P}_h^k$ to perform optimistic planning to approximate the joint optimal policy:

$$\nu_h^k(s) = \mathtt{CE}(\{\widehat{u}_h^{k,(i)}(s,\cdot)\}_{i\in[N]}),$$

$$\pi_h^k(s) = \arg\max_{b\in\mathcal{B}} \sum_{i=1}^{N} \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')\widehat{P}_h^k(\mathrm{d}s'|s,b),$$

where $\widehat{u}_h^k \in \mathcal{U}_h^k$ and $\widehat{P}_h^k \in \mathcal{P}_h^k$ are optimistic estimations.

Model-based Optimistic online Learning for MEE (MOLM)

MOLM algorithm analysis:

- Online regret for K episodes:

$$\mathrm{Regret}_{\mathsf{CE,SWM}}(K) = \sum_{k=1}^{K} \mathrm{SubOpt}(\pi^k, \nu^k).$$

- Sublinear regret of MOLM algorithm:

$$\mathrm{Regret}_{\mathsf{CE,SWM}}(K) \in \widetilde{\mathcal{O}}(H^2 N \sqrt{dK}),$$

  where H is the horizon, N is the number of agents, d is the eluder dimension of the function classes for general function approximations (Russo & Van Roy, 2013).

- Achieving $\widetilde{\mathcal{O}}(\sqrt{K})$-regret which is sublinear: MOLM efficiently finds the jointly optimal policy $(\pi^\star, \nu^\star)$ approximately.
- The key to achieve such regret is using the optimistic principle for exploration in uncertain environments.

Model-based Pessimistic offline Learning for MEE (MPLM)

MPLM algorithm design (two steps):

- **Model estimation step:** construct confidence sets $\mathcal{U}_h$ for utility functions and $\mathcal{P}_h$ for transition kernels using previously collected offline data only.
- **Pessimistic policy optimization step:** use $\mathcal{U}_h$ and $\mathcal{P}_h$ to perform pessimistic policy optimization to approximate the joint optimal policy:

$$\widehat{\nu}_h(s) = \texttt{CE}(\{\widehat{u}_h^{(i)}(s, \cdot)\}_{i \in [N]}),$$

$$(\widehat{\pi}, \widehat{P}) = \arg\max_{\pi \in \Pi} \min_{\widehat{P}: \{\widehat{P}_h \in \mathcal{P}_{h,\xi_2}, \forall h \in [H]\}} \sum_{i=1}^{N} \widehat{V}_{1,(\widehat{P},\widehat{u})}^{(\pi,\widehat{\nu}),(i)}(s_1),$$

where $\widehat{u}_h \in \mathcal{U}_h$ and $\widehat{P}_h \in \mathcal{P}_h$ are pessimistic estimations.

Model-based Pessimistic offline Learning for MEE (MPLM)

MPLM algorithm analysis:

- Offline suboptimality of MPLM algorithm:
$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}) \in \widetilde{\mathcal{O}}(H^2 N \sqrt{C^\star \iota / K}).$$
where H is the horizon, N is the number of agents, ι is the covering number of the function classes for general function approximations.
- $C^\star$ is the concentrability coefficient between data $\mathbb{D}$ and joint optimal policy $(\pi^\star, \nu^\star)$. Due to the use of pessimism principle, we only require the data to cover the joint optimal policy (partial coverage, rather than full coverage).
- Achieving $\widetilde{\mathcal{O}}(1/\sqrt{K})$-suboptimality: MPLM efficiently finds the jointly optimal policy $(\pi^\star, \nu^\star)$ approximately.

Thank You!