Pessimism meets VCG: Learning Dynamic Mechanism Design via Offline Reinforcement Learning

Boxiang Lyu[1] Zhaoran Wang[2] Mladen Kolar[1] Zhuoran Yang[3]

[1]UChicago Booth, [2]Northwestern University, [3]Yale University

Motivation

Dynamic mechanism design studies

- Allocation of goods in changing environments.
- Often formulate environments as MDPs.
- **Caveat:** often assumes environments are known a priori.

Can we recover a "good" dynamic mechanism with only access to a precollected dataset with offline RL, with no knowledge of the underlying MDP?

Preliminaries in MDP

An episodic MDP given by $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \{r_{i,h}\}_{i=0,h=1}^{n,H}\right)$.

- n agents, 1 seller.
- $\mathcal{S}$ state space, $\mathcal{A}$ action space, $\mathcal{P}$ transition kernel.
- $\forall i \in [n]$, $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0,1]$ agent i's reward function at step h. Seller's reward function is $r_{0,h} : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$.
- $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ seller's policy at step h. $V_h^{\pi}(\cdot; r)$ and $Q_h^{\pi}(\cdot; r)$ state- and action-value functions defined w.r.t. reward function r.

Dynamic Mechanism as an MDP

Interaction between buyer and seller.

- $h = 1$: WLOG environment starts at some $s_0 \in \mathcal{S}$.
- $h = 1, \ldots, H$:
  - Seller observes state s_h and takes action a_h, receiving reward $r_{0,h}(s_h, a_h)$.
  - Agents receive rewards $r_{i,h}(s_h, a_h)$ and report with a **potentially untruthful** reward function $\widetilde{r}_{i,h}(s_h, a_h)$.
  - Nature draws the next state $s' \sim \mathcal{P}_h(\cdot|s_h, a_h)$.
- $h = H$: Seller charges each agent i some price $p_i \in \mathbb{R}_+$.

The Markov VCG Mechanism

- Seller acts according to $\widetilde{\pi}^* = \operatorname{argmax}_\pi V_1^\pi(s_0; \sum_{i=0}^n \widetilde{r}_i)$.
- For $i \in [n]$, seller sets price p_i as follows

$$p_i = \max_\pi V_1^\pi(s_0; \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}^*}(s_0; \widetilde{R}_{-i}),$$

  where $\widetilde{R}_{-i} = \sum_{j \neq i} \widetilde{r}_i$.
- Intuition: p_i represents the "cost" of agent i joining the mechanism.

Mechanism Design Desiderata

Below we state, informally, three key mechanism design desiderata.

- Efficiency: the seller's policy maximizes the social welfare, i.e. the sum of rewards of all agents, when all agents report truthfully.
- Individual Rationality: the prices charged to the agents cannot exceed their rewards.
- Truthfulness: agents cannot increase their rewards by reporting untruthfully.

The Markov VCG mechanism satisfies all three simultaneously.
Can we show that we can learn a mechanism that satisfies all three approximately?

Estimating the Mechanism via Offline RL

Let $\mathcal{D} = \{(x_h^\tau, a_h^\tau, \{\widetilde{r}_{i,h}^\tau\}_{i=1}^n, x_{h+1}^\tau)\}_{h,\tau=1}^{H,K}$ be the dataset. We assume the entries are drawn i.i.d. from some distribution μ induced by some behavioral policy.

The intuition behind any algorithm that "learns" the Markov VCG mechanism can be summarized as follows.

- Step 1: use $\mathcal{D}$ to find some policy $\breve{\pi}$ such that $V_1^*(s_0; \widetilde{R}) - V_1^{\breve{\pi}}(s_0; \widetilde{R})$ is small.

- Step 2: for all i estimate the VCG price p_i as

$$\widehat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0),$$

where $G_{-i}^{(1)}(s_0)$ estimates $\max_\pi V_1^\pi(s_0; \widetilde{R}_{-i})$ and $G_{-i}^{(2)}(s_0)$ estimates $V_1^{\breve{\pi}}(s_0; \widetilde{R}_{-i})$.

Challenge: Estimating the VCG Price

Recall the VCG price estimate is given by

$$\widehat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0).$$

We highlight 3 challenges not found in prior works in offline RL.

1. Showing $\widehat{p}_i$ satisfies the mechanism design desiderata approximately.
2. Estimating $G_{-i}^{(1)}(s_0)$, which requires learning a fictitious policy that approximately maximizes $V_1^{\pi}(s_0; \widetilde{R}_{-i})$.
3. A combination of optimism and pessimism is needed for price estimation.

Policy Evaluation and Soft Policy Iteration

$B_{h,r}(f, \pi; \mathcal{D})$: empirical estimate for Bellman error under policy π at step h with respect to reward function r.

Policy evaluation procedure: solve the following problem

$$\operatorname{argmin}_{f \in \mathcal{F}} \pm f_1(s_0, \pi) + \lambda \sum_{h=1}^{H} B_{h,r}(f, \pi; \mathcal{D}),$$

where the first sign is $-$ if optimistic and $+$ if pessimistic.

Soft policy iteration: perform mirror descent-style updates

$$\widehat{\pi}_{h,r}^{(t+1)}(a|s) \propto \widehat{\pi}_{h,r}^{(t)}(a|s) \exp\left(\eta \widehat{Q}_{h,r}^{(t)}(s,a)\right),$$

where $\widehat{\pi}_{h,r}^{(t)}$, $\widehat{Q}_{h,r}^{(t)}$ can be optimistic or pessimistic, depending on the choice of policy evaluation procedure.

Summary of Results

When the dataset has sufficient coverage, the value functions are realizable by the function class $\mathcal{F}$, and the function class $\mathcal{F}$ is complete, with high probability

1. The social welfare suboptimality decays at a rate of $\mathcal{O}(K^{-2/3})$.
2. Seller's and agents' utility suboptimalities decay at a rate of $\mathcal{O}(K^{-2/3})$.
3. Agents' utilities are lower bounded by $-\mathcal{O}(K^{-2/3})$, i.e. the prices charged does not exceed their reward significantly.
4. Agents can gain at most $\mathcal{O}(K^{-2/3})$ from reporting untruthfully.

Particularly items 1 and 2 also requires truthful reporting from all agents.