The Grainger College of Engineering
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Online Learning for Dynamic Vickrey-Clarke-Groves Mechanism in Sequential Auctions under Unknown Environments

Vincent Leon leon18@illinois.edu
Rasoul Etesami etesami1@illinois.edu

ISE 2025 Student Conference

Introduction

Motivation

- Many real-world auctions (e.g., online ad allocation, allocation of CO2 emission licenses, wireless spectrum allocation, etc.) are **dynamic**.
- Bidders' values may change as the market environment **evolves**.
- The dynamics of the underlying environment is usually **unknown**.
- Existing learning-based VCG mechanisms use multi-armed bandits (MAB) and episodic Markov decision process (MDP) where the market resets. In practice, the market evolves **continuously**.

Our Goal and Contributions

- To extend the static VCG mechanism to dynamic auctions modeled as an **infinite-horizon average-reward MDP**.
- To design an online reinforcement learning (RL) algorithm for the seller to learn a dynamic mechanism that is **approximately efficient, truthful, and individually rational**.

Sequential Auctions Modeled as MDP

- 1 seller and n bidders
- State space $\mathcal{S}$: market conditions
- Action space $\mathcal{A}$: all possible allocations
- Transition kernel P: underlying dynamics
- Reward functions $\{r_i\}_{i=0}^n$: bidders' values
- Bidders submit bids $\{b_i\}_{i=1}^n$ to the seller
  - Truthful bidder: $b_i = r_i$
  - Untruthful bidder: otherwise
- The seller determines
  - Allocation policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
  - Price vector $p \triangleq (p_i)_{i=1}^n \in \mathbb{R}^n$

Technical Assumption

There exists some $\alpha > 0$ such that $P(s'|s,a) \geq \alpha$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

Offline Dynamic VCG Mechanism

... when the MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, \{r_i\}_{i=0}^n)$ is known.

Notation

Average social welfare (SW):

$$w(\pi) \triangleq J(\pi; R) \triangleq J(\pi; \textstyle\sum_{j=0}^n r_j) = \langle q^{P,\pi}, R\rangle$$

Bidder i's avg utility: $u_i(\pi, p) \triangleq J(\pi; r_i) - p_i$

Seller's avg utility: $u_0(\pi, p) \triangleq J(\pi; r_0) + \sum_{i=1}^n p_i$

Three Desiderata

- **Efficiency**:
  The mechanism maximizes the average SW when all bidders are truthful.
- **Truthfulness**:
  A bidder's average utility is maximized when she bids truthfully, regardless of the behavior of others.
- **Individual rationality**:
  A bidder's average utility is nonnegative when she bids truthfully, regardless of the behavior of others.

Infinite-horizon-MDP VCG Mechanism

- Allocation Policy π^*:
  $q^* \in \arg\max_{q \in \Delta(P)} \langle q, R\rangle \to \pi^* = \pi^{q^*}$
- Price Vector p^*:
  $p_i^* = \langle q_{-i}^* - q^*, R_{-i}\rangle$, where
  $q_{-i}^* \in \arg\max_{q\in\Delta(P)} \langle q, R_{-i}\rangle$, and $R_{-i} \triangleq \sum_{j\neq i} r_j$

THEOREM 1

*This mechanism is **efficient**, **truthful** and **individually rational**.*

Relaxed Desiderata for Online Learning

- **ϵ-Approximate efficiency**:
  $w(\pi^*) - \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T R^t\right] \leq \epsilon$ when all bidders are truthful.
- **Approximate truthfulness**:
  $\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T (\tilde{u}_i^t - u_i^t)\right] \leq 0$ when all other bidders adopt stationary bidding strategies (not necessarily truthful), where $\{u_i^t\}_{t=1}^T$ and $\{\tilde{u}_i^t\}_{t=1}^T$ are bidder i's realized utilities when she is truthful and untruthful, respectively.
- **Approximate individual rationality**:
  $\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T u_i^t\right] \geq 0$ when bidder i is truthful, regardless of the behavior of others.

Dual Formulation: Occupancy Measure

Given transition kernel P and stationary policy π:

$$q(s,a,s') \triangleq \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathbb{P}\{s^t = s, a^t = a, s^{t+1} = s'\}$$

$$\rho(s,a) \triangleq \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathbb{P}\{s^t = s, a^t = a\}$$

$\Delta(P)$, the set of all occupancy measures valid on P, is a **polytope**.

$\Delta \triangleq \cup_{P \text{ is valid}} \Delta(P)$ is a **polytope**.

A valid occupancy measure $q \in \Delta$ induces P and π:

$$P^q(s'|s,a) = \frac{q(s,a,s')}{\sum_{x\in\mathcal{S}} q(s,a,x)}$$

$$\pi^q(a|s) = \frac{\sum_{s'\in\mathcal{S}} q(s,a,s')}{\sum_{a'\in\mathcal{A}}\sum_{s'\in\mathcal{S}} q(s,a',s')}$$

Expected Average Reward and Occupancy Measure

$$J(\pi; r) \triangleq \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T r(s^t, a^t)\,\middle|\, s^1 = s\right]$$

$$= \langle q^{P,\pi}, r\rangle = \langle \rho^{P,\pi}, r\rangle$$

Online Learning for Dyn. VCG Mechanism

... when the MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, \{r_i\}_{i=0}^n)$ is unknown.

Difficulties in Online Learning for VCG Mech.

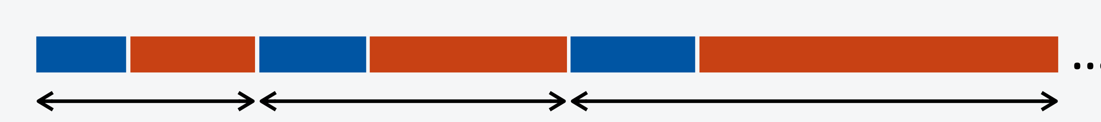
- Non-stationarity of MDP
- Learning and evaluation of policies not implemented
- Manipulation of seller's learning by untruthful bidders

Tackling the Difficulties

- Learning in episodes with increasing length
- Each episode divided into mixing and stationary phases
- Encouraged exploration using **shrunk occupancy measure polytope**

$$\Delta_\delta \triangleq \Delta \cap \left\{ q \in \mathbb{R}_+^{|\mathcal{A}||\mathcal{S}|^2} : \sum_{s'\in\mathcal{S}} q(s,a,s') \geq \delta, \forall s \in \mathcal{S}, a \in \mathcal{A}\right\}$$

Algorithm IHMDP-VCG



In episode k:

Mixing phase:

- For each round in the mixing phase:
  - Implement allocation policy $\pi^{[k]}$ induced by $\hat{q}^{[k]}$ and charge each bidder 0.
  - Collect reported rewards $\{r_i^t\}_{i=1}^n$ from the bidders.

Stationary phase:

- For each round in the stationary phase:
  - Implement allocation policy $\pi^{[k]}$ and charge each bidder $p_i^{[k]}$.
  - Collect reported rewards $\{r_i^t\}_{i=1}^n$ from the bidders.

Confidence sets update:

- Update confidence set for transition kernel $\mathcal{P}^{[k]}$:

$$\mathcal{P}^{[k]} \triangleq \mathcal{P}^{[k-1]} \cap \left\{ P \in \mathbb{R}_+^{|\mathcal{A}||\mathcal{S}|^2} : \left| P(s'|s,a) - \bar{P}^{[k]}(s'|s,a)\right| \leq \epsilon^{[k]}(s,a,s') \quad \forall (s,a,s')\right\}$$

- Update UCB and LCB for reward functions $\hat{r}_i^{[k]}$ and $\check{r}_i^{[k]}$.

Policy update:

- Update occupancy measure $\hat{q}^{[k+1]}$ by solving the following linear program (LP):
  $\hat{q}^{[k+1]} \in \arg\max_{q\in\Delta_\delta(\mathcal{P}^{[k]})} \left\langle q, \hat{R}^{[k]}\right\rangle$.
  (Remark: $\Delta_\delta(\mathcal{P}^{[k]})$ is a polytope.)
- For each bidder i, update payment $p_i^{[k+1]}$ by solving the following LP:
  $\hat{q}_{-i}^{[k+1]} \in \arg\max_{q\in\Delta_\delta(\mathcal{P}^{[k]})} \left\langle q, \hat{R}_{-i}^{[k]}\right\rangle$
  and using the following equation:
  $\hat{p}_i^{[k+1]} = \left\langle \hat{q}_{-i}^{[k+1]}, \hat{R}_{-i}^{[k]}\right\rangle - \left\langle \hat{q}^{[k+1]}, \check{R}_{-i}^{[k]}\right\rangle$.

THEOREM 2

*The algorithm IHMDP-VCG is **$\mathcal{O}(n\epsilon)$-approximately efficient, approximately truthful** and **approximately individually rational**.*