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

Vincent Leon

Department of Industrial & Enterprise Systems Engineering

University of Illinois Urbana-Champaign

Job Market Showcase, The 2025 INFORMS Annual Meeting, Atlanta, USA

Tuesday, October 28, 2025

Agenda

I. Introduction

II. Preliminaries—Infinite-horizon MDP and its Dual Formulation

III. Offline Dynamic VCG Mechanism—when the MDP is known

IV. Online Learning-based VCG Mechanism—when the MDP is unknown

V. Conclusion

I. Introduction

Vickrey-Clarke-Groves (VCG) Auctions

- Sealed-bid auction of multiple items
- Rational bidders submit bids that represent their values for the items
- The seller (or the mechanism) assigns the items and charges each bidder
- Three properties:
  - Efficient (socially optimal)
  - Truthful (incentive compatible)
  - Individually rational

Source: https://napga.org/its-time-for-the-2023-virtual-auction/

Motivation

- Many real-world auctions are dynamic.
  - Online ad allocation: [Branzei et al., 2023, Cramton and Kerr, 2002]
  - Allocation of CO2 emission licenses: [Balseiro and Gur, 2019, Golrezaei et al., 2019]
  - Wireless spectrum allocation: [Khaledi and Abouzeid, 2015, Milgrom, 2017]

Motivation

- Many real-world auctions are dynamic.
  - Online ad allocation: [Branzei et al., 2023, Cramton and Kerr, 2002]
  - Allocation of CO2 emission licenses: [Balseiro and Gur, 2019, Golrezaei et al., 2019]
  - Wireless spectrum allocation: [Khaledi and Abouzeid, 2015, Milgrom, 2017]
- Bidders' values may change as the market environment evolves.
- The dynamics of the underlying environment is usually unknown.
- Existing learning-based VCG mechanisms assume that the market resets.
  - Multi-armed bandits (MAB): [Kandasamy et al., 2023]
  - Episodic Markov decision process (MDP): [Lyu et al., 2022, Qiu et al., 2024]

In practice, the market evolves continuously.

Goal and Contributions

- To extend the static VCG mechanism to **sequential auction** 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**.

II. Preliminaries

Infinite-horizon MDP and its Dual Formulation

Dual Formulation: Occupancy Measure

In a unichain MDP:

- A transition kernel P and a stationary policy π define an occupancy measure:

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

  - Long-term probability that the state-action-next-state tuple (s,a,s') is visited
  - Dual variable of the MDP optimization problem

- A valid occupancy measure q induces a transition kernel P and a stationary policy π:

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

MDP Problem: From Primal to Dual

- $\Delta(P) \triangleq \{q^{P,\pi}$ for all stationary $\pi\}$ is a polynomial-sized polytope.
- $\Delta \triangleq \cup_P$ is valid $\Delta(P)$ is a polynomial-sized polytope.
- Expected average reward expressed using occupancy measure [Altman, 1999]

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

$$= \langle q^{P,\pi}, r\rangle \qquad \text{(Dual)}$$

- Dual of MDP optimization problem is a linear program (LP):

$$\max_{q\in\Delta(P)} \langle q, r\rangle$$

- From now on, the MDP problem will be written in its dual form.

III. Offline Dynamic VCG Mechanism

… when the MDP is known

Offline Sequential Auction Modeled as MDP

- Agents: 1 seller and n bidders
- Public information known to all agents:
  - State space $\mathcal{S}$: market conditions
  - Action space $\mathcal{A}$: all possible allocations
- Private information:
  - Each bidder $i \in [n]$ knows her own reward (value) function $r_i : \mathcal{S} \times \mathcal{A} \to [0,1]$.
  - The seller knows the transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$.

Offline Sequential Auction: Interaction Protocol

Before the sequential auction starts:

- Each bidder $i \in [n]$ submits her bids $b_i : \mathcal{S} \times \mathcal{A} \to [0,1]$ to the seller.
  - Truthful bidder: $b_i = r_i$
  - Untruthful bidder: otherwise
- The seller determines:
  - Allocation policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
  - Payment policy $p \triangleq (p_i)_{i=1}^n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$

After the sequential auction starts, the seller implements (π, p).

Three Desiderata for Offline Mechanism

- **Efficiency**:
  The mechanism maximizes the average social welfare 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.

Notation:

- Average social welfare: $w(\pi) \triangleq \langle q^{P,\pi}, \sum_{j=1}^{n} r_j \rangle$
- Bidder i's average utility: $u_i(\pi, p_i) \triangleq \langle q^{P,\pi}, r_i - p_i \rangle$

Infinite-horizon VCG Mechanism

Allocation Policy π^*

$$q^* \in \arg\max_{q \in \Delta(P)} \langle q, \sum_{j=1}^{n} r_j \rangle \quad \longrightarrow \quad \pi^* = \pi^{q^*}$$

Payment Policy p^*

$$p_i^*(s,a) = \max_{q \in \Delta(P)} \langle q, \sum_{j \neq i} r_j \rangle - \sum_{j \neq i} r_j(s,a) \quad \forall i, s, a$$

THEOREM 1

This dynamic mechanism is efficient, truthful and individually rational.

IV. Online Learning-based VCG Mechanism

… when the MDP is unknown

Online Sequential Auction Modeled as RL Problem

- Agents:
  - Learning agent: seller
  - Non-learning agents: bidders
- Public information known to all agents:
  - State space $\mathcal{S}$
  - Action space $\mathcal{A}$
- Unknown information:
  - The seller does not know the transition kernel P.
  - Each bidder $i \in [n]$ does not necessarily know her own reward function r_i.

Online Sequential Auction: Interaction Protocol

In each round t:

- The seller determines:
  - Allocation policy π^t
  - Payment policy $p^t \triangleq (p_i^t)_{i=1}^n$
- The seller:
  - Observes the state s^t
  - Chooses an allocation $a^t \sim \pi^t(\cdot \mid s^t)$
  - Charges $p_i^t(s^t, a^t)$ to each bidder $i \in [n]$
- Each bidder $i \in [n]$:
  - Receives a bandit feedback $r_i^t(s^t, a^t)$
  - Submits a bid $b_i^t \in \mathbb{R}$ for the next round
    (truthful bidder: $b_i^t = r_i^t(s^t, a^t)\ \forall t$; untruthful bidder: o.w.)

Relaxed Desiderata for Online Learning-based Mech. (1)

ϵ-Approximate efficiency:

$$w(\pi^*) - \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=0}^{n} r_j^t\right] \leq \epsilon$$

when all bidders are truthful.

Relaxed Desiderata for Online Learning-based Mech. (2)

Approximate truthfulness:

$$\lim_{T\to\infty} \sup \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} (\tilde{u}_i^t - u_i^t)\right] \leq 0$$

when all other bidders adopt stationary bidding strategies (not necessarily truthful), where

$\{\tilde{u}_i^t\}_{t=1}^T$: bidder i's realized utilities when she is untruthful,

$\{u_i^t\}_{t=1}^T$: bidder i's realized utilities when she is truthful.

Relaxed Desiderata for Online Learning-based Mech. (3)

Approximate individual rationality:

$$\liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} u_i^t\right] \geq 0$$

when bidder i is truthful, regardless of the behavior of others.

Recall: Offline Mechanism

Allocation Policy π^*

$$q^* \in \arg\max_{q \in \Delta(P)} \langle q, \sum_{j=1}^{n} r_j \rangle \quad \longrightarrow \quad \pi^* = \pi^{q^*}$$

Payment Policy p^*

$$p_i^*(s,a) = \max_{q \in \Delta(P)} \langle q, \sum_{j \neq i} r_j \rangle - \sum_{j \neq i} r_j(s,a) \quad \forall i, s, a$$

Naturally, we design an algorithm that learns P and $\{r_i\}_{i=1}^{n}$ and solves the LPs above iteratively.

What makes this problem more challenging than a single-agent RL problem?

Challenges and Solutions

Challenges:

1. Non-stationarity of MDP
2. Learning and evaluation of the policies not implemented
3. Manipulation of seller's learning outcome by untruthful bidders

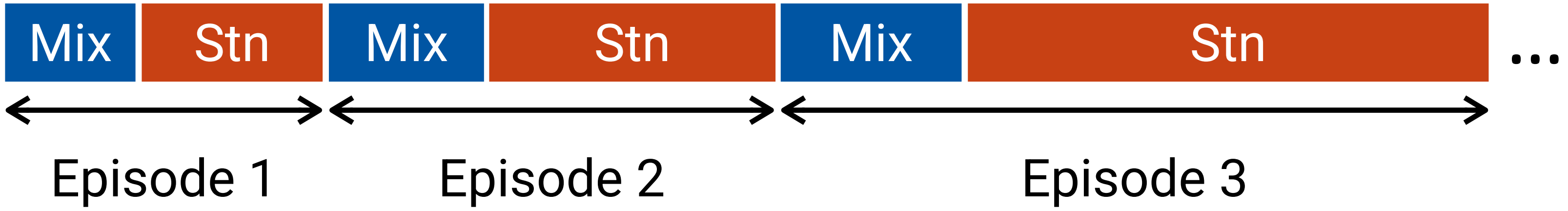
Solutions:

a. Learning in episodes with increasing length → 1

b. Each episode divided into mixing and stationary phases → 1, 2, & 3

c. Encouraged exploration by implementing stochastic policies only → 2 & 3
(“peeling off” the facets of the polytope that give deterministic policies → shrunk polytope)

Algorithm IHMDP-VCG



In each episode k:

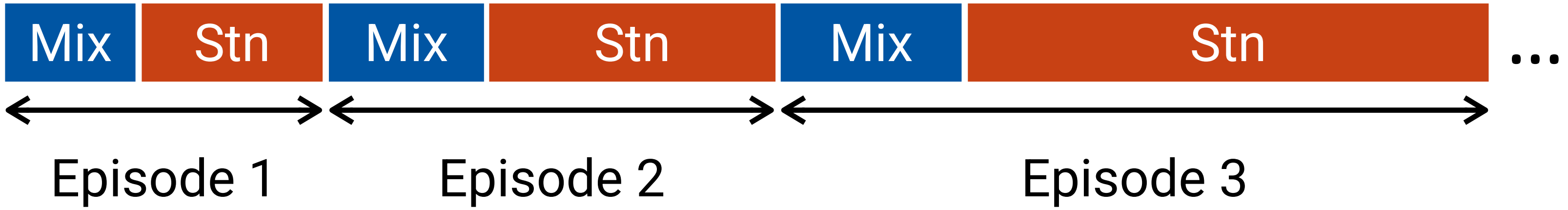
Mixing Phase:

- In each round:
  - Implement allocation policy $\pi^{[k]}$.
  - Charge each bidder 0.
  - Collect reported rewards $\{r_i^t\}_{i=1}^n$ from the bidders.

Stationary Phase:

- In each round:
  - Implement allocation policy $\pi^{[k]}$.
  - Charge each bidder $\hat{p}_i^{[k]}$.
  - Collect reported rewards $\{r_i^t\}_{i=1}^n$ from the bidders.

Algorithm IHMDP-VCG



At the end of episode k:

- Update confidence set for transition kernel $\mathscr{P}^{[k]}$.
- Update UCB and LCB for reward functions $\hat{r}_i^{[k]}$ and $\check{r}_i^{[k]}$.
- Update allocation policy:

$$\hat{q}^{[k+1]} \in \arg \max_{q \in \Delta_\delta(\mathscr{P}^{[k]})} \langle q, \sum_{j=0}^{n} \hat{r}_j^{[k]} \rangle \longrightarrow \pi^{[k+1]}$$

(Remark: $\Delta_\delta(\mathscr{P}^{[k]})$ is a shrunk polytope.)

- Update payment policy $\hat{p}^{[k+1]}$:

$$\hat{p}_i^{[k+1]}(s, a) = \max_{q \in \Delta_\delta(\mathscr{P}^{[k]})} \langle q, \sum_{j \neq i} \hat{r}_j^{[k]} \rangle - \sum_{j \neq i} \check{r}_j^{[k]}(s, a)$$

$$\forall i, s, a \, .$$

Main Results

THEOREM 2

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

V. Conclusion

Conclusion

- We have extended the static VCG mechanism to **dynamic sequential auction** modeled as an **infinite-horizon average-reward MDP**, preserving efficiency, truthfulness, and individual rationality.
- We have designed an online RL algorithm to learn a dynamic mechanism that achieves $\mathcal{O}(n\epsilon)$**-approximate efficiency, approximate truthfulness, and approximate individual rationality**.

References

Eitan Altman. 1999. *Constrained Markov decision processes: stochastic modeling.* Routledge. https://doi.org/10.1201/9781315140223

Santiago R Balseiro and Yonatan Gur. 2019. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science* 65, 9 (2019), 3952–3968.

Simina Branzei, Mahsa Derakhshan, Negin Golrezaei, and Yanjun Han. 2023. Learning and Collusion in Multi-unit Auctions. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 22191–22225.

Peter Cramton and Suzi Kerr. 2002. Tradeable carbon permit auctions: How and why to auction not grandfather. *Energy Policy* 30, 4 (2002), 333–345. https://doi.org/10.1016/S0301-4215(01)00100-8

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. 2019. Dynamic Incentive-Aware Learning: Robust Pricing in Contextual Auctions. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

Kirthevasan Kandasamy, Joseph E. Gonzalez, Michael I. Jordan, and Ion Stoica. 2023. VCG Mechanism Design with Unknown Agent Values under Stochastic Bandit Feedback. *Journal of Machine Learning Research* 24, 53 (2023), 1–45. http://jmlr.org/papers/v24/20-1226.html

Mehrdad Khaledi and Alhussein A. Abouzeid. 2015. Dynamic Spectrum Sharing Auction With Time-Evolving Channel Qualities. *IEEE Transactions on Wireless Communications* 14, 11 (2015), 5900–5912. https://doi.org/10.1109/TWC.2015.2443796

Boxiang Lyu, Zhaoran Wang, Mladen Kolar, and Zhuoran Yang. 2022. Pessimism meets VCG: Learning Dynamic Mechanism Design via Offline Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 162), Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 14601–14638. https://proceedings.mlr.press/v162/lyu22b.html

Shuang Qiu, Boxiang Lyu, Qinglin Meng, Zhaoran Wang, Zhuoran Yang, and Michael I. Jordan. 2024. Learning Dynamic Mechanisms in Unknown Environments: A Reinforcement Learning Approach. *Journal of Machine Learning Research* 25, 397 (2024), 1–73. http://jmlr.org/papers/v25/23-0159.html

Thank You

Questions?

Vincent Leon
leon18@illinois.edu
vin-leon.github.io


ILLINOIS