JOURNAL OF
EMERGING INVESTIGATORS

Article

Evaluating the performance of Q-learning-based AI in auctions

Moshi Liu[1], Jian Liu[1]

[1] Chadwick High School, Palos Verdes Peninsula, California

SUMMARY

In modern advertising platforms like Google Ads, advertisers set campaign goals and budgets, while artificial intelligence (AI) driven algorithms handle the actual bidding to maximize the advertiser's benefit. This study investigates whether AI-driven bidding strategies, developed through quality learning (Q-learning), align with classical auction theory. As a reinforcement learning algorithm, Q-learning learns by trial and error, improving actions based on rewards. In first-price auctions where the winner pays the highest bid, economic theory predicts that bidders should engage in bid shading, converging to 50% of their valuation in a two-bidder setting. In contrast, in second-price auctions where the winner pays the 2nd highest bid, bidding on 100% of their valuation is the dominant strategy. We hypothesized that Q-learning agent bidders will learn to bid around 50% of their valuation in first-price auctions and their full valuation in second-price auctions, aligning with theoretical predictions. Our results show that in first-price auctions, Q-learning bidders do not adjust their bids as theory predicts, instead stabilizing at 94.9% of their valuation. In second-price auctions, Q-learning agent bidders exhibit a near-truthful bidding pattern, converging to 98.5% of their valuation. Our analysis suggests that Q-learning does not adapt its strategy based on auction type. These findings highlight the limitations of reinforcement learning in capturing strategic reasoning, suggesting that current AI models struggle to develop auction-specific strategies without explicit guidance.

INTRODUCTION

As artificial intelligence (AI) technology gains more attention, an increasing number of advertising platforms now rely on automated online auctions. In Google, each user's search comes back with some advertisements (AD). The AD display opportunity triggers a real-time auction, creating millions of auctions every second (1). To handle this scale and speed, advertisers set high-level goals such as budgets or desired conversions, while AI-based algorithms turn these goals into bidding strategies to maximize the advertiser's benefit (1,2). As AI becomes more integrated into economic systems, it is important to examine whether classical auction theory still applies in AI-driven settings. Classical auction theory has thoroughly studied different types of auctions, deriving the optimal bidding strategies for bidders (3). However, AI-driven bidders using reinforcement learning techniques, like quality learning (Q-learning), introduce new dynamics that require further investigation (4,5).

In this study, we restrict our analysis to single-item, single-round auctions, where bidders submit one bid for a single item and the highest bidder wins. To better understand strategic behavior in such settings, we first review classical auction theory. Vickrey established the foundation of classical auction theory by analyzing optimal bidding strategies in first-price and second-price auctions (3). In a first-price auction, the highest bidder wins and pays the amount they bid, whereas in a second-price auction, the highest bid still wins but pays the second-highest bid (3). Classical theory predicts that bidders in first-price auctions will engage in bid shading, reducing their bids below their true valuation to maximize profit (3). This value is called equilibrium, which no bidder has an incentive to deviate from if their competitor is also following the same bidding strategy (3). Formally, in a first-price auction with N bidders, the equilibrium bidding strategy for bidder i is:

$$bid(v_i) = \frac{N-1}{N} \cdot v_i \qquad \textbf{(1)}$$

where v_i is the player i's private valuation of the item (in dollars), and $bid(v_i)$ is the corresponding equilibrium bid (3). In a two-bidder setting, where N = 2, the bidders should bid half of their valuation to maximize expected payoffs (3). In contrast, second-price auctions encourage truthful bidding, making it the best strategy for bidders to submit their true valuation regardless of competitors' actions (3).

More recent studies have incorporated reinforcement learning into auction mechanisms, demonstrating how AI enhances adaptability and performance in dynamic settings, where performance is defined as the ability to win auctions while keeping costs low (6-12). Banchio and Skrzypacz explored how repeated auctions help bidders improve their strategies through learning (13). Other researchers have shown that reinforcement learning can achieve strong performance in real-world applications, including microgrids, cognitive radio, and the energy market (6-11). While these studies investigating reinforcement learning show that AI can enhance bidding efficiency, they primarily focus on models designed to optimize performance (4-13). In contrast to this performance-driven perspective, our study explores a different question: can a simple, rule-based Q-learning agent independently evolve human-like strategic intelligence in auctions?

To explore this question, we examined whether Q-learning can develop optimal bidding strategies without built-in domain knowledge. Q-learning is a popular reinforcement learning framework to train the algorithm to learn optimal actions by interacting with an environment and receiving rewards (14). The Q-learning algorithm maintains a Q-table to store the expected reward for taking a specific action in each state and gradually improves its actions to achieve the maximum rewards (14). If the Q-learning agent autonomously converges to economic prediction, it suggests that Q-learning can evolve strategic intelligence through learning alone. If not, it may reflect limitations in how Q-learning adapts in competitive settings. Understanding these capabilities and constraints is essential for anticipating challenges in real-world AI deployment and improving future AI-driven decision-making systems.

We hypothesized that in two-bidder first-price auctions, a Q-learning agent would learn to engage in strategic bid shading, ultimately stabilizing at approximately 50% of their valuation, as predicted by economic theory. We also hypothesized that in second-price auctions, the Q-learning agent will learn to bid truthfully, stabilizing at 100% of their valuation, since truthful bidding is the best strategy in this setting. To test these hypotheses, we built a Q-learning-based simulation framework to model bidder strategies. Our results show that in first-price auctions, the Q-learning agent stabilized at 94.9% of their valuation, significantly deviating from the expected value of 50%. This implies that Q-learning prioritizes winning over profit maximization, failing to learn optimal bid shading. In second-price auctions, the Q-learning agent stabilized at 98.5%, which is close to the expected 100%. However, further analysis suggests that Q-learning agents rely on trial-and-error learning rather than strategic reasoning. Although the Q-learning agent adjusts its behavior based on past rewards, it does not demonstrate an understanding of auction structure or equilibrium strategies. While this helps the Q-learning agent make good bids in some cases, it also causes consistent mistakes in situations that need more strategic thinking, like first-price auctions. At the same time, it remains an open question whether these deviations stem from our specific reward design or from more fundamental limitations of Q-learning. Briefly raising this possibility highlights the need for future research to compare Q-learning with alternative reinforcement learning methods.

RESULTS

To study bidders' strategic behavior, we created a Q-learning-based simulation in which two bidders competed in 10,000 sequential auctions, defined as one simulation run. In this study, the terms 'simulation' and 'experiment' are used interchangeably to describe computational experiments. Each run generated a timeline of 10,000 auction outcomes. We repeated the simulation 1,000 times independently, averaging the winning bids at each time point to reduce random variation. All subsequent analyses are based on these averaged results.

In first-price auctions, our simulations converged to 94.9% of their true valuation, a significant deviation from the theoretical prediction of 50% (3). The convergence pattern suggests that Q-learning agents gradually increased their bids based on cumulative experiences (**Figure 1**).

In second-price auctions, convergence occurred at 98.5% of the true valuation, indicating a slight but consistent deviation from the theoretical expectation of 100% (**Figure 2**). To further illustrate this convergence pattern, we plotted the raw bid trajectory of a single Q-learning agent over time. The trajectory shows an initial phase of wide exploration followed by a gradual stabilization near the agent's true valuation (**Figure 3**).

We conducted one-sample *t*-tests with 95% confidence intervals to assess whether the winning bids had statistically converged in the final 1,000 auctions. In both first-price and second-price auctions, the test yielded a *p*-value of 0.999, indicating highly stable convergence across 1,000 auctions. These results validate the reliability of our findings (**Table 1**).

To assess the robustness of our findings, we extended the simulations by varying the Q-learning hyperparameters and by increasing the test scale to 150,000 auctions. Across all tested combinations, Q-learning agents in first-price auctions continued to converge to values far above the theoretical equilibrium of 50%, while in second-price auctions, convergence remained close to truthful bidding (**Table 2**).

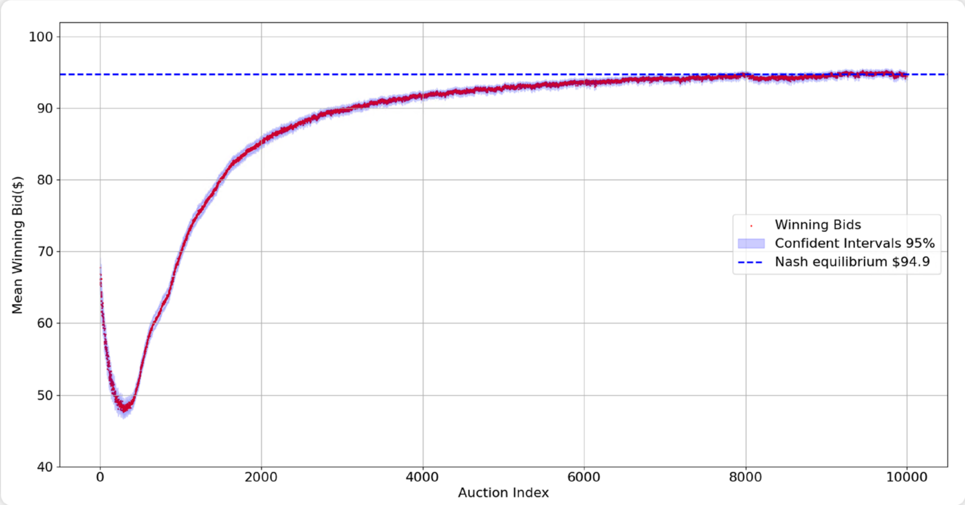

Figure 1: Winning bids across auction rounds under a first-price auction setting. Each red dot represents the mean winning bid for the *i*-th auction, calculated from 1,000 independent experimental runs over 10,000 auctions. The purple shaded region indicates the 95% confidence intervals derived from the standard deviation across these runs. The blue dashed line marks the observed convergence value of $94.9. A one-sample *t*-test was conducted from the last 1,000 auctions to evaluate convergence to the observed value of $94.9, and the results confirmed no significant deviation (*p* = 0.999).

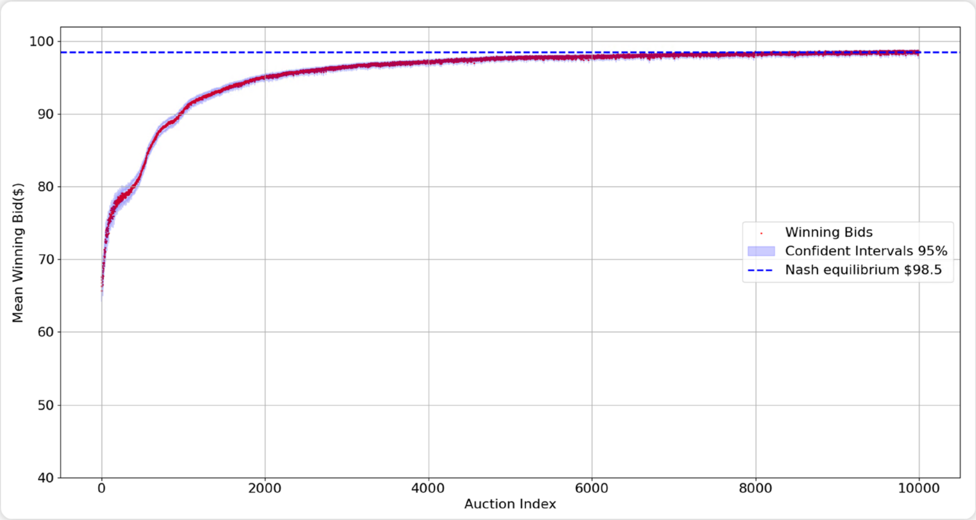

Figure 2: Winning bids across auction rounds under a second-price auction setting. Each red dot represents the mean winning bid for the *i*-th auction, calculated from 1,000 independent experimental runs over 10,000 auctions. The purple shaded region indicates the 95% confidence intervals derived from the standard deviation across these runs. The blue dashed line marks the observed convergence value of $98.5. A one-sample *t*-test was conducted from the last 1,000 auctions to evaluate convergence to the observed value of $98.5, and the results confirmed no significant deviation (p = 0.999).

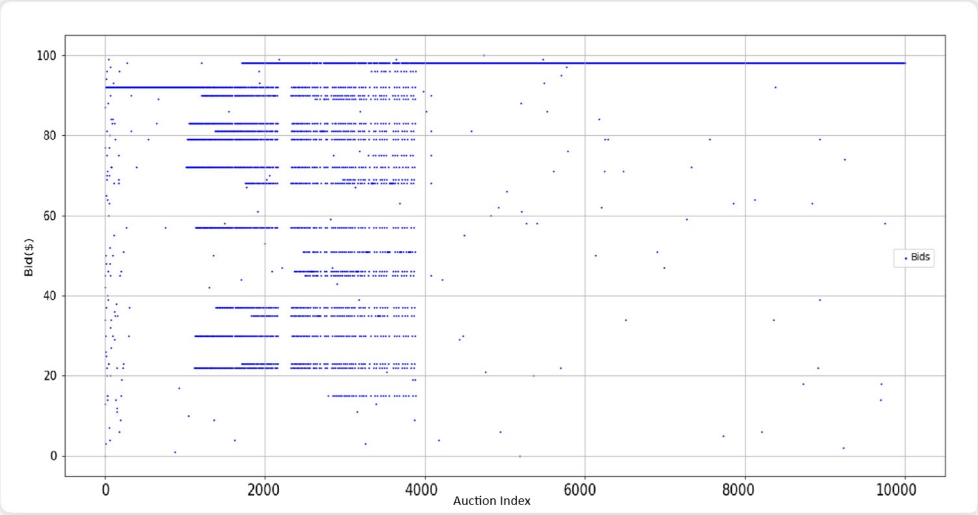

Figure 3: Raw bid trajectories in second-price auctions (α = 0.1, γ = 0.95). Each blue dot represents a bid submitted in an individual auction by a single Q-learning agent. The horizontal axis shows the auction index, and the vertical axis shows the bid amount. The scattered early-round bids reflect the Q-learning agent's trial-and-error exploration phase. And the concentration of bids near $100 in later rounds indicates convergence toward near-truthful bidding.

We also tested the multiple bidders' settings: in each simulation run, 100 bidders competed in 1,000 sequential auctions. This simulation was repeated 100 times independently, and the results were averaged across runs to improve robustness. The results showed that Q-learning agents converged to 99.1% of their valuation (p = 0.999) in first-price auction, and 100% of their valuation (p = 0.999) in second-price auction (**Table 3**).

DISCUSSION

Our experiments revealed that Q-learning performs well in second-price auctions but deviates significantly from equilibrium strategies in first-price settings. To better understand these behaviors, it is important to consider why Q-learning was chosen and how its characteristics relate to the auction environment.

Q-learning, a widely studied value-based reinforcement learning algorithm, was selected because of its simplicity and proven effectiveness in prior auction-related studies (6-11). Compared to policy-based or model-based methods, Q-learning offers an interpretable framework well-suited for environments with small, discrete state-action spaces (14). While other more advanced reinforcement learning algorithms, like Deep Q-learning or actor-critic, have been successfully applied to complex environments with high-dimensional or continuous state spaces, they introduce non-deterministic

α	γ	1st Price ($)	p-value(1st)	2nd Price ($)	p-value(2nd)
0.1	0.95	94.9	0.999	98.5	0.999

Table 1: Converged winning bid values and p-value for both first-price auctions and second-price auctions in the main experiments. Data comes from 1,000 independent experiments, with each experiment containing 10,000 auctions. The *t*-test is done based on the average bid over the final 1,000 auctions of each experiment.

convergence behavior (15-18). Since our primary goal is to investigate whether algorithm can learn strategic bidding through basic reward-driven learning mechanisms, tabular Q-learning provides a transparent and controlled framework (14).

We used a basic function as the reward function for Q-learning agent. This function was deliberately chosen to test the agent's ability to learn strategy purely from outcome-based feedback, without embedding prior knowledge. Although more sophisticated or context-aware reward structures could provide stronger learning signals and help agents distinguish between auction formats, those algorithms would blur the line between learned intelligence and engineered incentives.

In a first-price auction, with only two bidders participating, the equilibrium is for each bidder to bid half of their valuation (3). In theory, this equilibrium should emerge as bidders refine their strategies over time (3). However, our results show a significant deviation, with the Q-learning agent converging to 94.9% instead of the predicted 50% equilibrium (**Figure 1**). One possible reason for this discrepancy is that our Q-learning agents lack rational equilibrium assumptions. Auction theory assumes that all bidders rationally follow the equilibrium strategy (3). However, agents in our model use Q-learning, which relies on trial-and-error exploration rather than explicit knowledge of equilibrium strategies. Unlike human bidders who might analytically compute the optimal bidding function, Q-learning agents adjust their behavior based on past rewards, potentially leading to different convergence points (14).

Learning equilibrium strategies purely through Q-learning, without encoding them into the algorithm, remains a challenge. It requires human-like intelligence and involves reasoning about the competitor's behavior besides its own actions. At present, the most popular AI technologies are based on neural networks, which primarily perform pattern matching from existing data rather than genuine logical reasoning (19). While neural networks have achieved impressive results, they lack true reasoning abilities. Whether they can lead to human-like intelligence remains uncertain, and researchers continue to explore alternative approaches (19).

Another factor is the limited competition in the two-bidder setting. The classical 50% shading strategy emerges as an equilibrium only when bidders expect competition from one opponent (3). In a first-price auction with *N* bidders, as *N* grows large toward infinity, bidders should bid close to their true value (**Equation 1**). However, our Q-learning agent does not recognize the difference between competing against one bidder vs. multiple bidders. To test whether this pattern holds in larger auctions, we conducted an additional experiment with 100 bidders. The results showed that Q-learning agents converged to 99.1% of their valuation, which is only slightly higher than the 94.9% observed in the two-bidder setting (**Table 3**). This reinforces the idea that our Q-learning agent treats auctions as if the number of competitors were infinite.

α	γ	1st Price($)	p-value(1st)	2nd Price($)	p-value(2nd)
0.05	0.50	94.6	0.99	99.5	0.99
0.10	0.50	95.0	0.99	99.4	0.99
0.30	0.50	95.4	0.99	99.2	0.99
0.05	0.80	93.2	0.99	99.1	0.99
0.10	0.80	93.9	0.99	99.1	0.99
0.30	0.80	95.0	1.0	98.9	0.99
0.05	0.99	92.7	0.99	97.5	0.99
0.10	0.99	91.3	0.99	97.5	0.99
0.30	0.99	86.4	0.99	94.7	0.99
0.10	**0.95**	**92.1**	**0.99**	**98.0**	**0.99**

Table 2: Converged winning bid values and p-value for both first-price and second-price auctions for extended experiments with different α and γ. Each row corresponds to a unique combination of α and γ to examine their impact on bid convergence. Data comes from 1,000 independent experiments, with each of the experiments containing 150,000 auctions. The *t*-test is done based on the average bid over the final 10,000 auctions of each experiment. The last row (in bold) is the factors we chose for the main experiments.

Unlike first-price auctions, where bidders must strategize since no dominant strategy exists, second-price auctions have a different theoretical outcome. Here, a bidder's best move is independent of the opponent's bid, eliminating the need for strategic adjustments (3).

Our results show that Q-learning agents in second-price auctions did not fully converge to 100% of their valuation but instead stabilized at 98.5%. While this is very close to the expected equilibrium, it remains slightly lower. To investigate whether the Q-learning agent's near-truthful bidding behavior arises from strategic awareness or from simple reward maximization, we examined the raw bid trajectory of a single Q-learning agent over 10,000 auction rounds (**Figure 3**). The plot reveals a broad initial exploration followed by a gradual and steady increase in bid values. Importantly, there is no indication of an abrupt behavioral transition that would suggest the Q-learning agent recognized the dominant strategy associated with this auction setting. Instead, the Q-learning agent's bid values shift upward in response to higher rewards obtained from winning with higher bids. This trajectory supports the conclusion that the Q-learning agent's learning process is purely reinforcement-driven, and the convergence to near-truthful bidding is coincidental rather than reflective of auction-specific strategic reasoning.

While these findings offer valuable insights into the learning behavior of Q-learning agents in auction environments, they also highlight several limitations. Addressing these limitations will be important for extending the generality and robustness of our research. One limitation concerns the length of the training rounds. Our experiments used 10,000 auctions for training. The winning bid curve indicates that bidding behavior is unstable during the initial rounds, suggesting that shorter runs may not provide sufficient experience for strategy convergence (**Figures 1 and 2**). To test the stability of our chosen horizon, we extended the simulations to 150,000 auctions and observed that the agent's behavior remained stable after 10,000 rounds, supporting the robustness of our default training length (**Table 2**). However, we did not test shorter horizons in this research, and we acknowledge that a

α	γ	1st Price ($)	p-value(1st)	2nd Price ($)	p-value(2nd)
0.1	0.95	99.1	0.999	100	0.999

Table 3: Converged winning bid values and *p*-value for both first-price auctions and second-price auctions with 100 bidders. Data comes from 100 independent experiments, with each experiment containing 1,000 auctions. The *t*-test is done based on the average bid over the final 100 auctions of each experiment.

sensitivity analysis would be valuable in future work to better understand the convergence behavior under different training durations. Another important limitation in our research is the simplified reward function used in Q-learning. In the second-price auction, where the optimal strategy is straightforward, our Q-learning with basic intelligence performs well, as simple reinforcement suffices to arrive at the correct behavior. However, in the first-price auction, where strategic bid shading requires human-like intelligence to reason about opponents' actions, our Q-learning framework struggles. This suggests that our simple reward design may not provide enough guidance for strategic behavior, or that the Q-learning algorithm may struggle in general to account for opponents and make truly strategic decisions.

Our observation raises a broader question: is the observed performance gap specific to our Q-learning setup, or does it reflect challenges intrinsic to value-based reinforcement learning, or even to reinforcement learning more broadly? Unlike our simplified model, commercial bidding systems (e.g., Google Ads) integrate historical data, contextual signals, and market dynamics to enable more adaptive strategic bidding behaviors (1,2,20). These systems go beyond simple state-action learning and incorporate opponent modeling, uncertainty estimation, and budget-aware optimization (1,2,20). To address these open questions, future research should compare Q-learning with other reinforcement learning approaches, such as State–Action–Reward–State–Action (SARSA, an on-policy value-based method), Deep Q-learning (function approximation), and policy-gradient or actor-critic methods (direct policy optimization) (15-18). Such benchmarking could help isolate whether the observed performance gap stems from the reward design, the limitations of tabular value iteration, or broader weaknesses in reinforcement learning strategies in auction settings.

MATERIALS AND METHODS

The Model

Q-learning utilizes a learning process involving a continuous balance between exploration and exploitation (14). Exploration occurs when the bidder experiments with different bid values, selecting them randomly within a range, typically [*0*, v_i], where v_i represents how much the bidder values the item. In contrast, exploitation happens when the bidder leverages past experiences to make more informed decisions, selecting bids that have previously resulted in higher rewards rather than continuing to experiment (14).

In each auction, the Q-learning agent decides whether to explore or exploit by generating a random value between 0 and 1, then compare this value to the exploration-exploitation trade-off parameter ε (14). If the random value is less than ε, the Q-learning agent selects exploration; otherwise, it opts for exploitation. The ε starts at 1.0 to encourage exploration and gradually decreases over time following a decay strategy to balance exploration and exploitation (14). The decay is applied using the formula: *ε = max (0.01, ε * 0.99)*.

In the Q-learning algorithm, two important hyperparameters control the learning process: the learning rate (α) and the discount factor (γ). The parameter α determines the weight given to new information (14). The discount factor γ influences how much Q-learning agent values future rewards over immediate ones (14). In our model, α = 0.1 and γ = 0.95 are chosen for the main testing as they provide a balance between learning stability and adaptation speed. In addition, other parameter combinations were explored, which are detailed in the tuning section.

The Q-table is the core of our Q-learning agent's decision-making process. In our setup, the bid price is defined as the Q-table state. After each round, the Q-learning agent updates its Q-table based on the received reward, using the Bellman equation (**Equation 2**). The Q-table stores states and their corresponding expected rewards for taking specific actions. Over time, as the bidder explores more, the Q-table becomes increasingly accurate, allowing for better predictions and decision-making.

$$Q(s,a) \leftarrow \mathrm{Q}(\mathrm{s},\mathrm{a}) + \alpha\,[\mathrm{r} + \gamma \max\left(\mathrm{Q}(\mathrm{s}',\mathrm{a}')\right) - \mathrm{Q}(\mathrm{s},\mathrm{a})] \qquad \textbf{(2)}$$

where *Q(s, a)* is the current Q-value for state *s* and action *a, α* is the learning rate, *r* is the reward in state *s* for taking action *a*, *γ* is the discount factor, and *max(Q(s', a'))* is the max future reward for all possible actions *a'* in the next state *s'*.

The reward function used in our model is defined as follows (**Equation 3**):

$$reward(bid) = \begin{cases} valuation - payment, & if\ win \\ 0, & if\ lose \end{cases} \qquad \textbf{(3)}$$

This reward function was applied in all simulations.

Statistical Analysis

In our experiments, two bidders compete for a single item, with both bidders' valuations fixed at $100 per auction. Each experiment consists of *N* = 10,000 sequential auctions, where each auction takes place at a discrete time step *i (0 ≤ i < N). M* = 1,000 experiments were run independently and the average winning bid values were used for analysis to reduce bias in the data. At each time *i*, corresponding to an auction, the average winning bid was computed across all experiments (**FIgure 4**). The average winning bid at time *i (*0 ≤ *i* < *N)* is given by:

$$avg\_winning\_bid[i] = \frac{1}{\mathrm{M}} \sum_{j=1}^{M} \left(winning\_bid_j[i]\right) \qquad \textbf{(4)}$$

where $winning\_bid_j[i]$ is the winning bid at time *i* in experiment *j*.

Only the last 1,000 auctions were considered to determine the final convergence value of the bidding curve. The early auction rounds corresponding to the learning phase were excluded from the calculation. A one-sample *t*-test was conducted to determine whether the final bid converges to the mean of the last 1,000 auctions. The null hypothesis (H_0) states that the mean winning bid from the last 1,000 auctions is equal to the converged value, while the alternative hypothesis (H_1) proposes that the mean winning bid is significantly different from the converged value. If the *p*-value

https://doi.org/10.59720/25-073

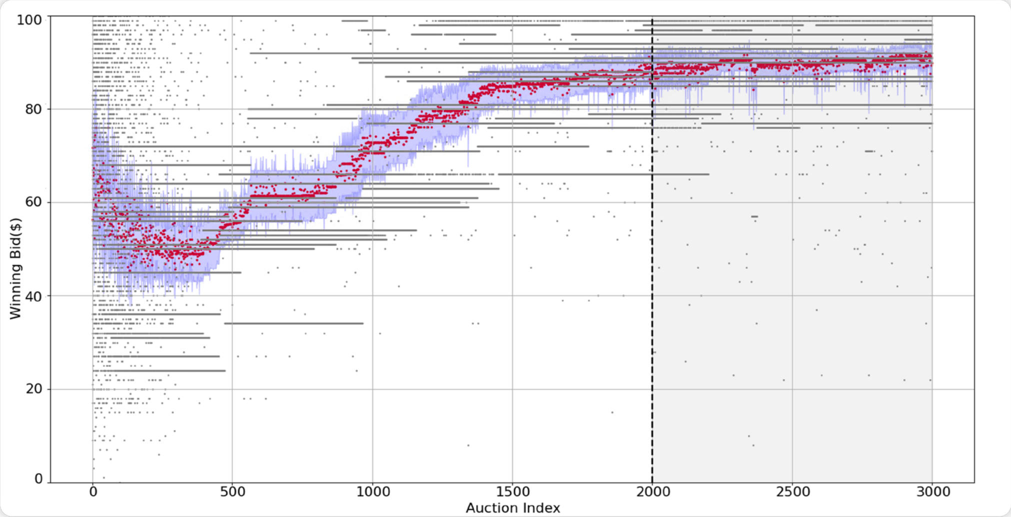

Figure 4: Visualization of winning bid computation, confidence intervals, and convergence value determination. The gray dots represent raw winning bid values from individual auctions for all M experiments, while the red dots indicate the mean winning bid at each time *i*, computed by averaging all M raw values at the corresponding x-coordinate. The purple shaded region represents the 95% confidence interval of this mean winning bid value. The final convergence value is derived from the red average winning bids in the rightmost gray shaded region, where statistical stability is confirmed using a one-sample *t*-test.

from the *t*-test is greater than 0.05 ($p \geq 0.05$), it fails to reject the null hypothesis, indicating that the bids have statistically converged. Conversely, if $p < 0.05$, it rejects H_0, suggesting that the bidding values have not yet stabilized.

Model Tuning

In Q-learning, the selection of hyperparameters plays a crucial role in an agent's ability to learn optimal strategies (14). Proper tuning of these parameters is essential to balance learning efficiency and bid convergence. To analyze their effect on bid convergence, three representative values were selected for each parameter and all possible combinations were tested in our experiments. For the learning rate (α), 0.05, 0.10, and 0.30 were chosen to represent slow, moderate, and fast learning rates (21). For the discount factor (γ), 0.50, 0.80, and 0.99 were tested to examine short-term, medium-term, and long-term reward considerations (21). To better demonstrate the stability of our results, each experiment was extended to include N = 150,000 auctions and expanded the convergence window to the last 10,000 auctions, further confirming that the observed convergence patterns remain consistent. Despite variations in specific parameter settings, all experiments yielded results consistent with the core findings of our main study (α =0.1, γ = 0.95) (**Table 2**).

Software and Libraries

All experiments were implemented in Python 3.10.11, and the code is available at https://github.com/AllenThePenguin/AuctionMLResearch/. The open-source library NumPy 2.2.3 was used for array-based computations, SciPy 1.15.1 was used for statistical analysis, and Matplotlib 3.10.0 was used for rendering the graphs (22-24).

Received: February 18, 2025
Accepted: October 5, 2025
Published: November 9, 2025

REFERENCES

1. Cai, Han, *et al*. "Real-Time Bidding by Reinforcement Learning in Display Advertising." *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2 February 2017, pp. 661–670. https://doi.org/10.1145/3018661.3018702.
2. Varian, Hal R. "Position Auctions." *International Journal of Industrial Organization*, vol. 25, no. 6, 16 November 2007, pp. 1163–1178. https://doi.org/10.1016/j.ijindorg.2006.10.002.
3. Vickrey, William. "Counterspeculation, auctions, and competitive sealed tenders." *The Journal of Finance*, vol. 16, no. 1, March 1961, pp. 8–37. https://doi.org/10.2307/2977633.
4. Khezr, Peyman, *et al*. "Artificial Intelligence for Multi-Unit Auction Design." *arXiv Preprint arXiv:2404.15633*, 8 August 2024. https://doi.org/10.48550/arXiv.2404.15633.
5. Rodriguez-Fernandez, Jaime., et al. "Context Aware Q-Learning-Based Model for Decision Support in the Negotiation of Energy Contracts." *International Journal of Electrical Power & Energy Systems*, vol. 104, January 2019, pp. 489–501. https://doi.org/10.1016/j.ijepes.2018.06.050.
6. Avval, Akram Esmaeili, *et al*. "The Comparison of Pricing Methods in the Carbon Auction Market via Multi-Agent Q-Learning." *RAIRO-Operations Research*, vol. 55, no. 3, May-June 2021, pp. 1767–1785. https://doi.org/10.1051/ro/2021065.
7. Wang, Ning, *et al.* "A Q-Cube Framework of Reinforcement Learning Algorithm for Continuous Double Auction among Microgrids." *Energies*, vol. 12, no. 15, 26 July 2019, p. 2891. https://doi.org/10.3390/en12152891.
8. Abbass, Waseem, *et al*. "Channel Allocation to GAA Users Using Double Deep Recurrent Q-Learning Based on Double Auction Method." *IEEE Access*, vol. 11, 20 October 2023, pp. 117321–117340. https://doi.

org/10.1109/ACCESS.2023.3326432.

9. Chen, Zhe, *et al*. “Q-Learning Based Bidding Algorithm for Spectrum Auction in Cognitive Radio.” *2011 Proceedings of IEEE Southeastcon*, 21 April 2011, pp. 409–412. https://doi.org/10.1109/SECON.2011.5752976.
10. Jiang, Jun, *et al*. “Deep Reinforcement Learning-Based Bidding Strategies for Prosumers Trading in Double Auction-Based Transactive Energy Market.” *arXiv Preprint arXiv:2502.15774*, 16 February 2025. https://doi.org/10.48550/arXiv.2502.15774.
11. Thirunavukkarasu, Gokul Sidarth, *et al.* “Advancing Transactive Energy Market Management Using Community Microgrid Emulator That Supports OpenADR and Q-Learning Based Auction Model.” *2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG)*, 02 February 2024, pp. 1–6. https://doi.org/10.1109/etfg55873.2023.10407269.
12. Asker, John, *et al*. "Artificial intelligence, algorithm design, and pricing." *AEA Papers and Proceedings*, vol. 112, May 2022, pp. 452–456. https://doi.org/10.1257/pandp.20221059.
13. Banchio, Martino, and Andrzej Skrzypacz. "Artificial intelligence and auction design." *Proceedings of the 23rd ACM Conference on Economics and Computation*, 13th July 2022, pp. 30–31. https://doi.org/10.1145/3490486.3538244.
14. Watkins, Christopher JCH, *et al.* "Q-learning." *Machine learning*, vol. 8, no. 3, May 1992, pp. 279–292. https://doi.org/10.1007/BF00992698.
15. Mnih, Volodymyr., *et al.* “Human-level control through deep reinforcement learning.” *Nature*, vol. 518, no. 7540, 25 February 2015, pp. 529–533. https://doi.org/10.1038/nature14236.
16. Wang, Yin-Hao, *et al.* “Backward Q-Learning: The Combination of Sarsa Algorithm and Q-Learning.” *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, October 2013, pp. 2184–2193. https://doi.org/10.1016/j.engappai.2013.06.016.
17. Peters, Jan, *et al.* “Reinforcement Learning of Motor Skills with Policy Gradients.” *Neural Networks*, vol. 21, no. 4, May 2008, pp. 682–97. https://doi.org/10.1016/j.neunet.2008.02.003.
18. Konda, Vijay, *et al*. “Actor-Critic Algorithms.” *Advances in Neural Information Processing Systems*, vol. 12, 1999. https://doi.org/10.1137/S0363012901385691.
19. Mirzadeh, Iman, *et al.* "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models." *arXiv preprint arXiv:2410.05229*, 7 October 2024. https://doi.org/10.48550/arXiv.2410.05229.
20. Perlich, Claudia, *et al.* “Machine Learning for Targeted Display Advertising: Transfer Learning in Action.” *Machine Learning*, vol. 95, no. 1, 30 May 2013, pp. 103–27. https://doi.org/10.1007/s10994-013-5375-2.
21. Schneckenreither, Manuel, *et al.* “Average Reward Adjusted Discounted Reinforcement Learning.” *Neural Computing and Applications*, 18 January 2025, pp. 1–32. https://doi.org/10.1007/s00521-024-10620-5.
22. Harris, Charles R., *et al.* “Array programming with NumPy.” *Nature*, vol. 585, no. 7825, 16 September 2020, pp. 357–362. https://doi.org/10.1038/s41586-020-2649-2.
23. Virtanen, Pauli, *et al.* “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods*, vol. 17, no. 3, 03 February 2020, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2.
24. Hunter, John D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering*, vol. 9, no. 3, May-June 2007, pp. 90–95. https://doi.org/10.1109/MCSE.2007.55.