Localized Epsilon-Greedy Q-learning and Price Competition in Amazon-Like Buybox Markets

Keywords: collusion, algorithmic pricing, platforms.

Hashem Amireh

March 2, 2025

Abstract: *This paper examines the performance of Q-learning algorithms in a "winner-take-all" pricing environment, such as the Amazon Buybox. Motivated by the increasing use of algorithms for pricing and the potential for collusion, this paper makes two contributions. First, it investigates whether collusive pricing trends observed in previous literature persist even in a setting designed to promote price competition, like the Buybox. Second, it proposes a refinement to the standard epsilon-greedy Q-learning algorithm called Localized Epsilon-Greedy Q-learning, where price exploration is focused around the current best price estimate. Through simulations, I demonstrate that the Buybox mechanism does not prevent collusive pricing and that localized exploration can lead to faster convergence to the monopoly price and higher profits for firms compared to the traditional epsilon-greedy approach.*

1 Introduction

Algorithmic pricing, the use of machine learning algorithms to dynamically adjust prices, is rapidly transforming online markets. As the market share of e-commerce continues to rise, firms increasingly rely on these algorithms to maximize profits. However, this trend raises concerns about potential anti-competitive outcomes. When multiple firms employ algorithms in strategic contexts like online ad auctions, pricing on e-commerce platforms, and online rentals, there's a risk of collusion and artificially inflated prices.

This paper investigates algorithmic pricing in a specific "winner-take-all" setting, exemplified by the Amazon Buybox. While the Buybox doesn't literally capture all sales, the winning firm typically secures a significant majority, making it crucial for firms to compete aggressively for this privileged position. One might expect that a mechanism like the Buybox, which inherently favors lower prices, would prevent collusive outcomes and promote price competition. However, I show throiugh simulations that this is not necessarily the case.

I focus on epsilon-greedy Q-learning, a specific type of reinforcement learning algorithm. To understand this algorithm, it's helpful to consider its place within the broader landscape of machine learning algorithms.

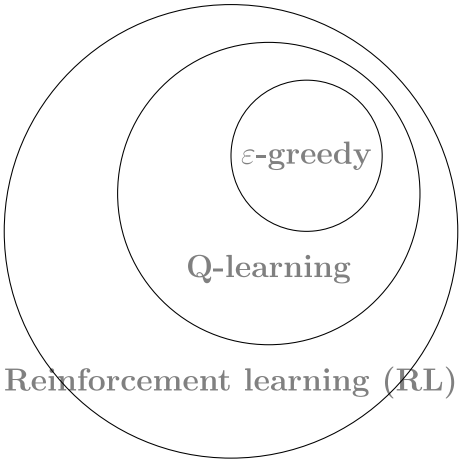


Figure 1: Types of Algorithms

Reinforcement Learning (RL) is a broad class of algorithms focuses on learning through trial and error. An RL agent interacts with an environment, takes actions, receives feedback in the form of rewards, and adjusts its strategy to maximize cumulative rewards over time.

Q-learning is specific type of model-free RL algorithm. "Model-free" means that the algorithm doesn't need to know the underlying dynamics of the environment; it learns directly from experience. Q-learning works by estimating the value (the "Q-value") of taking a particular action in a given state. Following each action, the algorithm updates the Q-value for that action.

Epsilon-Greedy Q-learning is a refinement that specifies a tradeoff between exploration and exploitation. With probability ε, the algorithm explores by taking a random action.

With probability $1 - \varepsilon$, it exploits its current knowledge by choosing the action with the highest Q-value.

This paper builds upon epsilon-greedy Q-learning and makes two key contributions: First, I investigate whether the collusive pricing trends observed in previous literature persist even in a Buybox environment designed to encourage price competition. I investigate whether the Buybox mechanism effectively prevents algorithmic collusion, or whether Q-learning algorithms still learn to collude and maintain supra-competitive prices. Through simulations, I show that price collusion is still possible in this setting as the algorithm considers dynamic aspect of returns.

Second, I propose a further refinement of the Epsilon-Greedy Q-learning algorithm which I call "Localized Epsilon-Greedy Q-learning". This addresses a key limitation of traditional Epsilon-Greedy Q-learning: its inability to consider the inherent relationships between prices. By focusing exploration on prices near the current best estimate, I expect an improvement in learning efficiency and firm performance. I investigate whether Localized Epsilon-Greedy Q-learning leads to faster learning, quicker convergence to the monopoly price, and higher profits for firms compared to the standard epsilon-greedy approach in a winner-take-all market. Through simulations, I show that the localized algorithm can indeed outperform the traditional algorithm but only under a certain set of relevant parameters.

This paper is structured as follows: Section 2 provides a literature review. Section 3 present the model that will be uses as a basis for the simulation. Section 4 outlines the simulation procedure, parameters, and results. Section 5 provides a discussion of the results. Section 6 concludes.

2 Literature Review

The study of algorithmic pricing and its potential for collusion has gained significant traction in recent years. Several recent papers have explored the ability of algorithms to learn to collude, using both theoretical and experimental approaches. Calvano et al. (2020) were among the first to demonstrate that epsilon-greedy Q-learning algorithms can lead to tacit collusion in a Bertrand pricing game, although the likelihood of collusion is sensitive to simulation parameters. They show experimentally that even simple Q-learning algorithms can learn to charge supracompetitive prices. The algorithms achieve this by coordinating on punishment schemes that discourage price cuts, and these schemes are learned without any explicit communication or instructions to collude.

Brown and MacKay (2021) document that some online retailers use algorithms that allow for more frequent price changes and automated responses to price changes by rivals. They develop a model of price competition that incorporates increased pricing frequency and short-run commitment through the use of algorithms. The authors show that algorithmic competition can generate price dispersion, increase price levels, and exacerbate the price effects of mergers.

Banchio and Mantegazza (2022) develop a tractable model for studying strategic interactions between learning algorithms, and uncover a mechanism responsible for the emergence of algorithmic collusion. The authors observe that algorithms periodically coordinate on actions that are more profitable than static Nash equilibria. This novel collusive channel relies on an endogenous statistical linkage in the algorithms' estimates which the authors call "spontaneous coupling". They also explore the design of strategy-proof mechanisms in the presence of algorithmic players, emphasizing the role of feedback on counterfactuals to reduce collusion.

Klein (2021) examines the collusive capacity of Q-learning in a simulated sequential-pricing duopoly setting. The author finds that competing Q-learning algorithms can often coordinate on collusive equilibria when the set of discrete prices is limited. As the set of discrete prices increases, the algorithms increasingly converge to a stable supra-competitive asymmetric cycle.

Possnig (2023) presents an analytical characterization of the long-run policies learned by algorithms that interact repeatedly. The author considers a repeated Cournot game of quantity competition, in which learning the stage game Nash equilibrium serves as a non-collusive benchmark. The author gives necessary and sufficient conditions for this Nash equilibrium not to be learned, which are requirements on the state variables algorithms use to determine their actions, and on the stage game. When algorithms determine actions based only on the past period's price, the Nash equilibrium can be learned. However, agents may condition their actions on richer types of information beyond the past period's price. In that case, the author gives sufficient conditions such that the policies converge with positive probability to a collusive equilibrium, while never converging to the Nash equilibrium.

Johnson, Rhodes, and Wildenbeest (2021) use both economic theory and extensive experiments on artificial intelligence pricing algorithms to demonstrate that platforms may indeed be able to design rules that increase competition on their marketplaces. Relatively simple rules may suffice when firms (or their algorithms) behave competitively, but more subtle ones-which condition on past behavior and treat firms asymmetrically-may be required when there is a risk of collusion.

While these studies provide valuable insights, they do not specifically address the winner-take-all dynamics prevalent in platforms like Amazon. Moreover, they often employ standard Q-learning algorithms that overlook the ordinal and cardinal relationships between prices. My work aims to fill this gap by examining collusion in a Buybox setting and introducing Localized Epsilon-Greedy Q-learning.

3 Model

3.1 Primitives

I consider a market with N firms, M potential consumers, and T discrete time periods. In each period t, firms simultaneously choose prices $\mathbf{p_t} = \{p_{1t}, p_{2t}, ..., p_{Nt}\}$ from a set of possible prices. The platform assigns the Buybox to one firm based on their price relative to the previous winner, as described in Section 3.3. Let $BB_{jt} \in \{0, 1\}$ denote whether firm j "has the buybox" in period t and $\mathbf{BB_t} = \{BB_{1t}, BB_{2t}, ..., BB_{Nt}\}$ denotes the set. Let $p_t^{BB_\tau}$ denote the price in period τ for the firm that won the buybox in period τ. More formally, if $BB_{jt} = 1$ then $p_t^{BB_t} := p_{jt}$

3.2 Firm (algorithm)

firms employ an epsilon-greedy Q-learning algorithm to choose prices. With probability ε, they explore by choosing a random price. With probability $1 - \varepsilon$, they exploit by selecting the price with the highest Q-value. Q-values are updated using the Bellman equation:

$$Q'(A) = (1 - \alpha)Q(A) + \alpha(R_t + \lambda \max_a Q(a))$$

where α is the learning rate, λ is the discount factor, and R_t is the observed profit in period t.

My proposed Localized Epsilon-Greedy Q-learning modifies the exploration step. Instead of choosing a completely random price, the algorithm samples a price from a truncated normal distribution (truncated at the price range) centered around the exploitation price (the one with the highest Q-value). The variance of this distribution is denoted σ_p. This localized exploration allows the algorithm to focus on prices near the current best estimate, potentially leading to faster learning and better performance.

3.3 Platform

Once the firms choose their prices, the platform then chooses the the buybox winner. However, in order to incentivize price competition, the platform can a set a threshold determining how much a firm must undercut the previous buybox winner by in order to "steal" the buybox from them. Let $\delta \in [0, 1]$ denote a parameter that determines how much a firm must undercut the previous buybox winner by to win the buybox. Further, let $p_{jt} = \min[\mathbf{p_t}]$ (i.e.

firm j has the lowest price in period t).

- If $t = 1$:

  Platform assigns $BB_{jt} = 1$ and $BB_{kt} = 0$ $\forall$k $\neq j$.

- If $t > 1$:

  The platform assigns $BB_{jt} = 1$ to firm j if:

  $p_{jt} < (1 - \delta)p_t^{BB_t}$ -or- $p_{jt} < (1 - \delta)p_t^{BB_{t-1}}$

  Otherwise, firm assigns $BB_{jt} = BB_{jt-1}$

3.4 Demand

Consumer demand is modeled through an indirect utility function:

$$u_{ijt} = \gamma_0 + \gamma_{1i}p_{jt} + \gamma_{2i}BB_{jt} + \nu_{ijt}$$

where γ_{1i} and γ_{2i} are consumer-specific parameters drawn from normal distributions s.t. $\gamma_{1i} \sim N(\bar{\gamma_1}, \sigma_{\gamma_1})$ and $\gamma_{2i} \sim N(\bar{\gamma_2}, \sigma_{\gamma_2})$, and ν_{ijt} is a Type I Extreme Value error term. Given the Type I Extreme Value distribution of the error term, the market share of firm j in period t can be expressed as:

$$s_{jt} \approx \iint \frac{exp(V_{ijt})}{1 + \sum_{k=1}^{N} exp(V_{ikt})} dF(\gamma_{1i})dF(\gamma_{2i})$$

where $V_{ijt} = \gamma_0 + \gamma_{1i}p_{jt} + \gamma_{2i}BB_{jt}$ represents the deterministic component of the utility function.

4 Simulation

To evaluate the performance of Localized Epsilon-Greedy Q-learning and investigate its implications in a winner-take-all market, I conduct a series of simulations. These simulations model the dynamic interactions between sellers, consumers, and the platform's Buybox mechanism. The simulation environment consists of two sellers who repeatedly engage in price competition, each using an Epsilon-Greedy Q-learning algorithm to choose prices. They face a large number of consumers who make purchase decisions based on a utility function that incorporates price, Buybox ownership, and individual preferences. The platform assigns the Buybox to one of the sellers in each period based on their prices and an undercut threshold (δ).

The simulation unfolds in a series of discrete time periods. First, I initialize the model parameters, including those governing consumer demand, seller behavior (learning rate, exploration rate, etc.), and the platform's Buybox assignment rule. Each seller begins with a Q-table initialized to zero for all possible prices. In each period, sellers use their Q-learning algorithms to determine prices. They either exploit their current knowledge by choosing the price with the highest Q-value or explore by choosing a random price (standard epsilon-greedy) or a price from a distribution around the current best price (localized epsilon-greedy). The platform then assigns the Buybox to one of the sellers based on their chosen prices and the undercut threshold.

Once the Buybox is assigned, consumers make purchase decisions based on their utility functions, leading to market shares for each seller. Sellers then realize profits based on their market shares and chosen prices. Finally, sellers update their Q-tables based on the observed profits, using the Bellman equation to learn and improve their pricing strategies over time. This sequence of price determination, Buybox assignment, market share calculation, and Q-value updates is repeated for a specified number of periods ($T = 1000$).

I compare the performance of the standard epsilon-greedy algorithm and the Localized Epsilon-Greedy algorithm across various values of σ_p (the standard deviation of the price distribution in the localized algorithm) and (the platform's undercut threshold). To assess performance, I track how quickly the algorithms converge to the monopoly price, which represents the collusive outcome in this setting, and I compare the cumulative profits earned by sellers under each algorithm.

The simulation results are presented in the Appendix and summarized in a heatmap (Figure 2). The heatmap shows the difference in total cumulative profits between the localized and regular algorithms for various combinations of σ_p and .

4.1 Simulation Results

I conducted simulations with the following parameters:

- **Demand parameters:** $M = 1000$, $\gamma_0 = 100$, $\bar{\gamma_1} = -5$, $\sigma_{\gamma_1} = 1.5$, $\bar{\gamma_2} = 30$, $\sigma_{\gamma_2} = 10$
- **Firm parameters:** $N = 2$, $c = 10$, $\alpha = 0.9$, $\varepsilon = 0.8$, $\lambda = 0.7$, $p \in [10, 40]$ with 0.10 increments.

I compared the performance of the regular and localized epsilon-greedy algorithms across various values of σ_p (localized variance) and δ (platform undercut threshold).

The appendix shows how the Q-values updates over time for traditional epsilon-greedy (blue) and localized epsilon-greedy for a given firm. It is clear that the localized algorithm converges to the area around the monopoly price (vertical green line) more quickly. Snapshots of the Q-values over time illustrate how the localized algorithm concentrates its learning around the optimal price more quickly than the regular algorithm.

This shows that regardless with our without the localized refinement, even with the Buybox mechanism, Q-learning algorithms can still learn to collude and maintain prices above competitive levels.

However, localized Epsilon-Greedy Q-learning leads to faster convergence to the monopoly price and generally higher cumulative profits for firms, particularly for moderate values of σ_p and δ.

Now if we allow σ_p and to vary, we see that the algorithm only performs better unda certain set of parameters. This is shown in Figure 2) where blue is the area at which the localize is more proftable while the red is the area where it is less profitable.

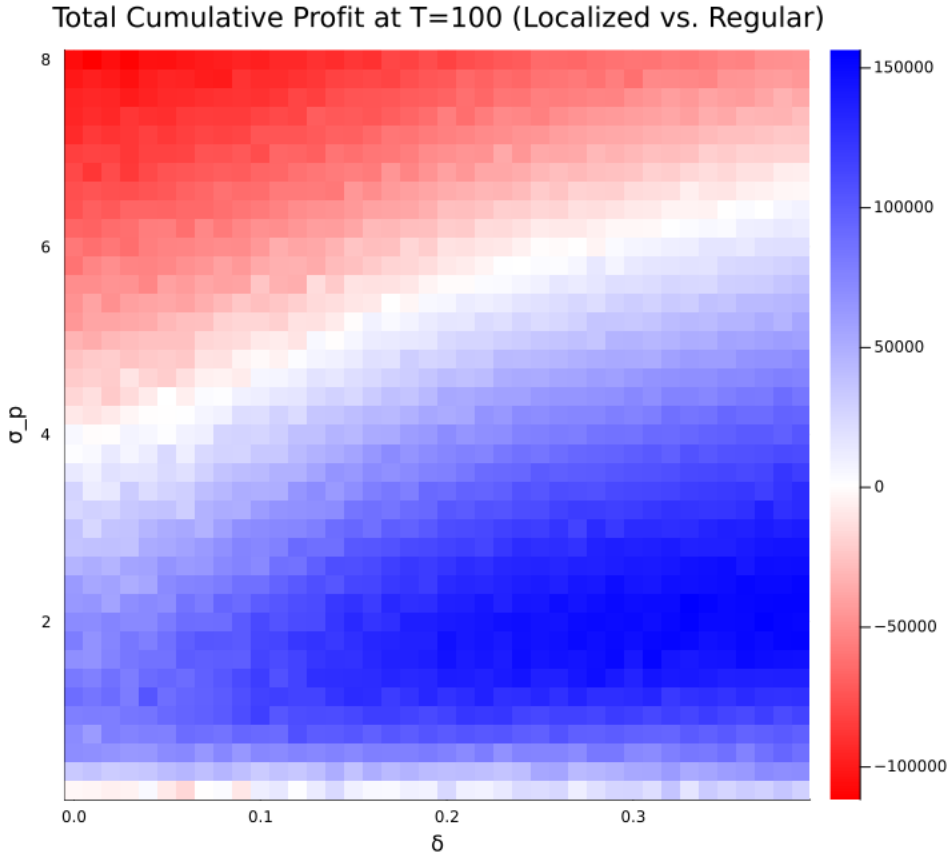


Figure 2: Difference in cumulative profit for both firms (1000 runs)

5 Discussion

My findings have important implications for both firms and platforms. For firms, Localized Epsilon-Greedy Q-learning offers a more efficient way to learn optimal pricing strategies in competitive markets. By focusing exploration on relevant price ranges, firms can potentially reach higher profits faster.

Platforms, on the other hand, should be aware of the potential for algorithmic collusion, even with simple learning algorithms and mechanisms designed to promote lower prices. The choice of platform parameters, such as the undercut threshold (δ), can significantly influence the competitive dynamics and outcomes.

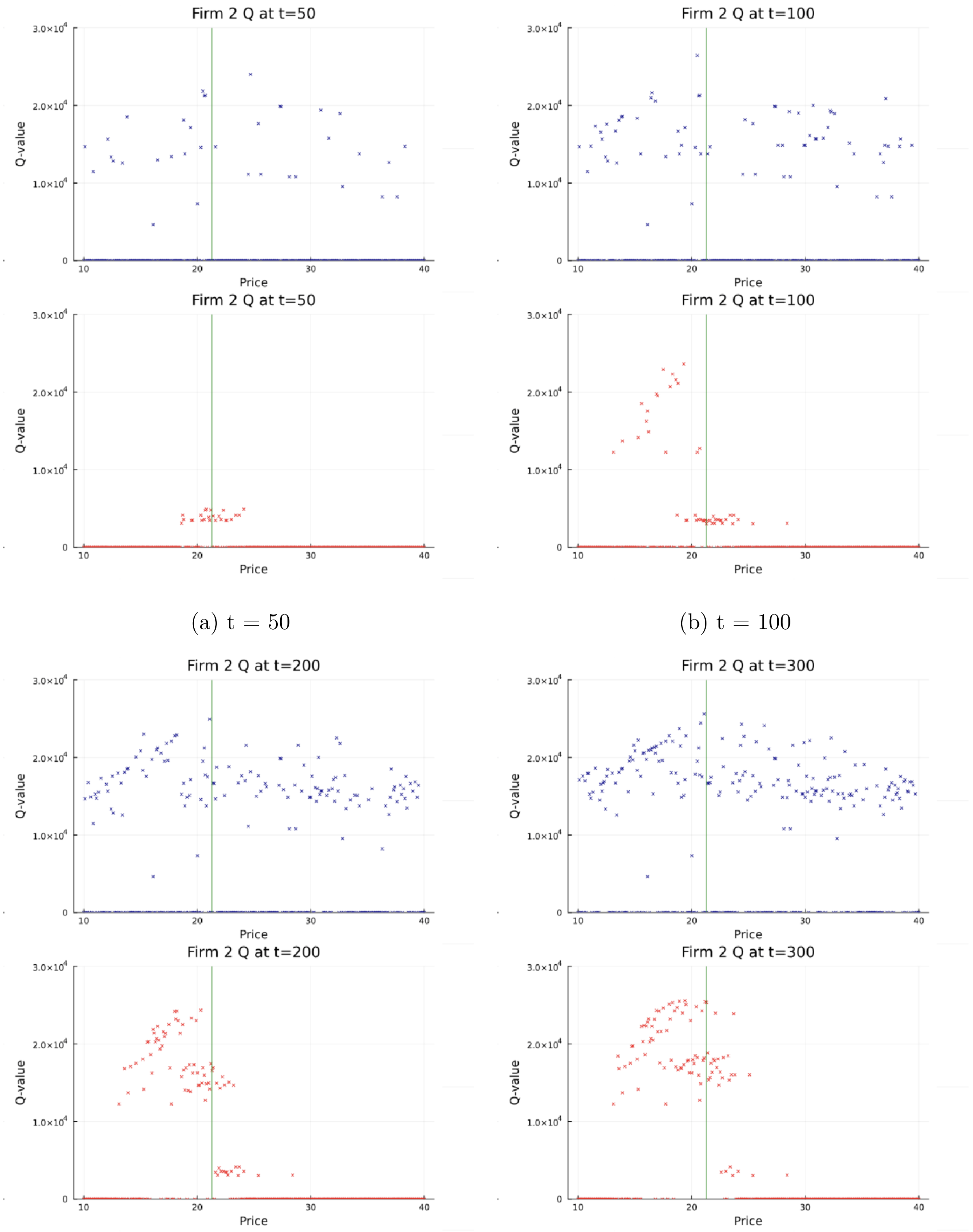
My model has limitations. I assume a relatively simple demand function and a stylized Buybox mechanism. Future research could explore more complex scenarios with heterogeneous consumers, dynamic demand, and more sophisticated platform algorithms. Additionally, further investigation into the optimal choice of σ_p and its potential decay over time is warranted.

While the simulations provide compelling evidence for the advantages of Localized Epsilon-Greedy Q-learning and highlight the potential for collusion even in a Buybox environment, establishing these findings rigorously through formal theorems presents a significant challenge. The complex interplay of dynamic learning, strategic interactions between sellers, and the platform's Buybox mechanism creates a complex system that is difficult to analyze analytically. For example, proving convergence to the monopoly price would involve demonstrating that the Q-learning algorithms, through repeated interactions and updates, eventually settle on a pricing strategy that sustains supra-competitive prices. Similarly, formally establishing the profitability of localized exploration would require characterizing the optimal range of σ_p values and demonstrating its superiority over the standard epsilon-greedy approach across different market conditions. While these theoretical explorations are beyond the scope of this paper, the simulation results serve as a valuable starting point for future research aimed at developing a deeper theoretical understanding of algorithmic pricing in winner-take-all markets.

6 Conclusion

This paper makes two important contributions to the literature on algorithmic pricing. First, it demonstrates that collusive pricing patterns can emerge even in winner-take-all markets like the Amazon Buybox, which are intended to foster price competition. Second, it proposes Localized Epsilon-Greedy Q-learning, a refinement to the standard epsilon-greedy approach that incorporates price relationships into the exploration process. My simulations show that this localized exploration leads to faster learning and better performance for firms.

Appendix A: Evolution of Q-values Over Time



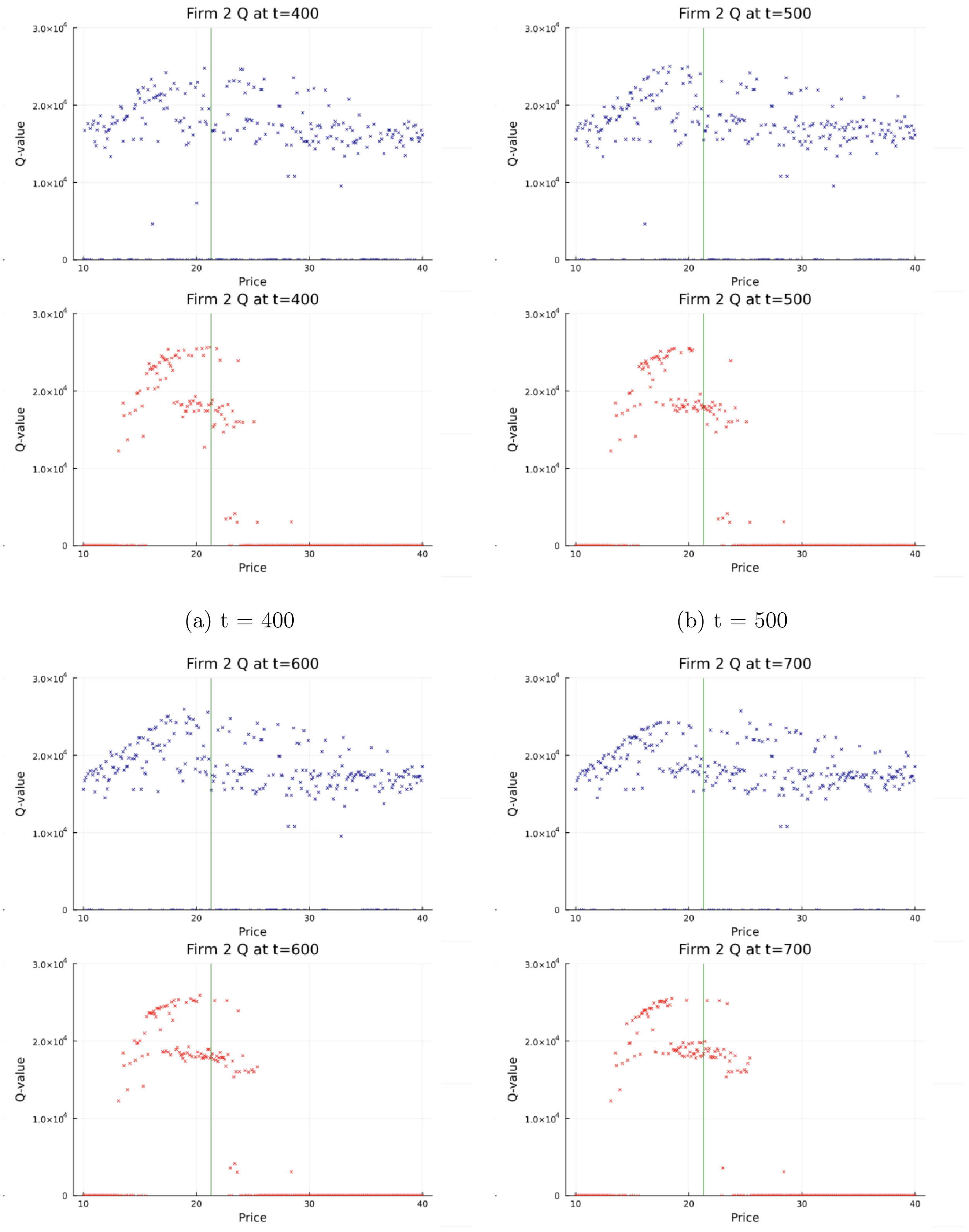
(a) t = 50

(b) t = 100

(c) t = 200

(d) t =300

Figure A1: Evolution of Q-values Over Time for a Given Firm



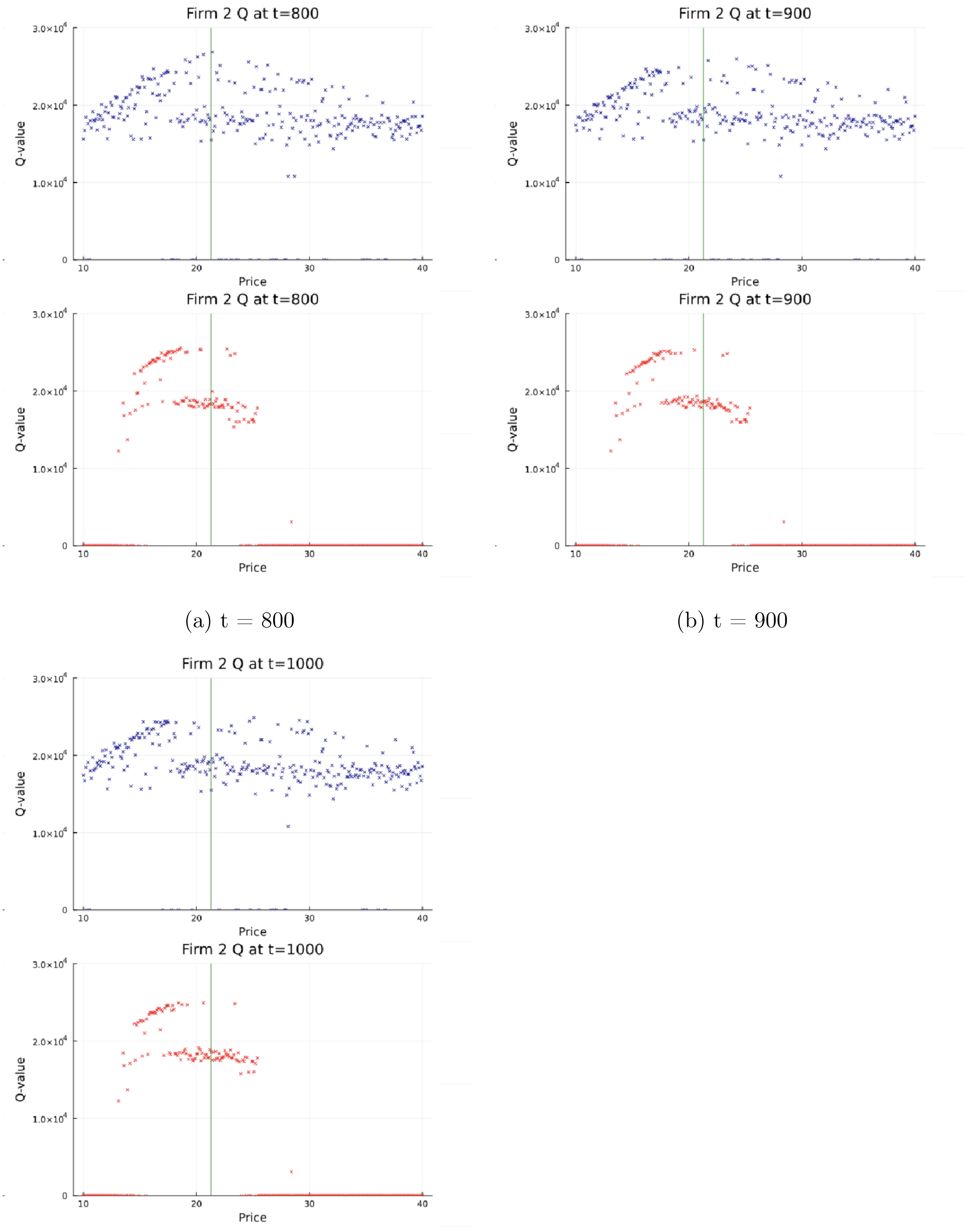
(a) t = 400

(b) t = 500

(c) t = 600

(d) t =700

Figure A2: Evolution of Q-values Over Time for a Given Firm



(a) t = 800

(b) t = 900

(c) t = 1000

Figure A3: Evolution of Q-values Over Time for a Given Firm

References

Banchio, Martino and Giacomo Mantegazza (2022). “Artificial intelligence and spontaneous collusion”. In: *arXiv preprint arXiv:2202.05946*.

Brown, Zach Y and Alexander MacKay (2021). “Competition in pricing algorithms”. In.

Calvano, Emilio et al. (2020). “Artificial intelligence, algorithmic pricing, and collusion”. In: *American Economic Review* 110.10, pp. 3267–97.

Johnson, Justin, Andrew Rhodes, and Matthijs Wildenbeest (2021). “Platform design when sellers use pricing algorithms”. In.

Klein, Timo (2021). “Autonomous algorithmic collusion: Q-learning under sequential pricing”. In: *The RAND Journal of Economics* 52.3, pp. 538–558.

Possnig, Clemens (2023). “Reinforcement Learning and Collusion”. In: *Available at SSRN 4506587*.