Learning Collusion in Episodic, Inventory-Constrained Markets

Paul Friedrich
University of Zurich,
ETH AI Center
Zurich, Switzerland
paul.friedrich@uzh.ch

Barna Pásztor
ETH AI Center,
ETH Zurich
Zurich, Switzerland
barna.pasztor@ai.ethz.ch

Giorgia Ramponi
University of Zurich,
ETH AI Center
Zurich, Switzerland
giorgia.ramponi@uzh.ch

ABSTRACT

Pricing algorithms have demonstrated the capability to learn tacit collusion that is largely unaddressed by current regulations. Their increasing use in markets, including oligopolistic industries with a history of collusion, calls for closer examination by competition authorities. In this paper, we extend the study of tacit collusion in learning algorithms from basic pricing games to more complex markets characterized by perishable goods withfi xed supply and sell-by dates, such as airline tickets, perishables, and hotel rooms. We formalize collusion within this framework and introduce a metric based on price levels under both the competitive (Nash) equilibrium and collusive (monopolistic) optimum. Since no analytical expressions for these price levels exist, we propose an efficient computational approach to derive them. Through experiments, we demonstrate that deep reinforcement learning agents can learn to collude in this more complex domain. Additionally, we analyze the underlying mechanisms and structures of the collusive strategies these agents adopt.[1]

KEYWORDS

Multi-Agent Reinforcement Learning; Game Theory; Pricing; Collusion; Bertrand Competition; Episodic Markov Game

ACM Reference Format:
Paul Friedrich, Barna Pásztor, and Giorgia Ramponi. 2025. Learning Collusion in Episodic, Inventory-Constrained Markets. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 24 pages.

1 INTRODUCTION

Algorithms are increasingly replacing humans in pricing decisions, offering improved revenue management and handling of complex dynamics in large-scale markets such as retail and airline ticketing. These algorithms, whether programmed or self-learning, can engage in tacit collusion charging *supra-competitive* prices (i.e., above the competitive level) or limiting production without explicit agreements. For example, algorithmic pricing in Germany led to a 38% increase in fuel retailer margins after adoption [7]. Our study is primarily motivated by airline revenue management (ARM), a market with $800 billion in annual revenue and thin profit margins. Airlines have already been under regulatory scrutiny [26] due to evidence of tacit collusion even before the introduction of algorithmic pricing [14] but the current trend of moving towards algorithmic pricing [35, 44] could lead to further cases.

Tacit collusion is maintained without explicit communication or agreement between sellers, therefore it eludes detection and often falls outside the scope of current competition laws. These concerns and potential negative effects on social welfare have been recognized both by regulators [17, 23, 41] and scholars [10, 15, 31]. To develop comprehensive legislation on algorithmic pricing, a thorough understanding of the factors that influence the emergence of collusive strategies is required under assumptions that align with real markets [19].

Previous research has already shown that *reinforcement learning (RL)* algorithms can engage in tacit collusion in pricing games with infinite time-horizon [6, 20, 34, 40]. However, most markets follow some form of periodicity, e.g., seasonality in retail orfi scal years for public companies, which breaks the continuity of the interactions between sellers. In the markets of perishable goods, hotels, or tickets, the markets only persist until the given sell-by dates and sellers are aware of thefi nite nature of competition. Importantly, in the previously investigated infinite time-horizon settings the collusive equilibrium is maintained via punishment strategies, e.g., grim-trigger, but it is not an equilibrium in thefi nite-horizon case. This is because these strategies are only credible if sufficient time remains for the punishment to offset short-term gains from deviating from collusion. In thefi nite-horizon setting, such punishments become unmaintainable as the sell-by date approaches. However, RL algorithms show the potential to learn collusion through their memory over several episodes interacting against the same opponents. Additionally, infi nite time-horizon markets supplies are often predetermined and limited, therefore, pricing strategies have to consider additional constraints and anticipate future demand to avoid expiring inventory while maximizing total profit. Both aspects are crucial in many real-world markets. For example, airlines selling tickets between two cities on a certain day have tofi ll their planes' capacity before departure. However, selling tickets too quickly could lead to a missed opportunity to sell tickets closer to departure time to less price-sensitive consumers, while selling tickets too slowly could result in empty seats. The added complexity offi nite time horizon and inventory constraints results in more complex strategies and interactions between pricing algorithms; therefore, previous results do not immediately hold and further investigation is necessary to develop comprehensive collusion mitigation approaches.

In this work, we aim to contribute to these efforts by extending the analysis of tacit collusion between pricing algorithms to *episodic markets with inventory constraints.*

[1]Code and data available at: https://github.com/pfriedric/EpisodicCollusion.
Preprint including appendix available at: https://arxiv.org/abs/2410.18871

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

In particular, in Section 2, we give an overview of related literature. In Section 3, we define the episodic,fi nite-horizon pricing problem with inventory constraints as a Markov game, inspired by Airline Revenue Management (ARM), and formalize both competitive (Nash) and collusive (monopolistic) equilibrium strategies. Building on these, we define a measure that quantifies collusion in an observed episode. Notably, our definitions are on the space of pricing strategies instead of price levels at a certain point in time which is the standard in the infinite-horizon setting. This is a significant change in the analysis and a challenge in episodic markets compared to the infinite-horizon case. In Section 4, we discuss how our model'sfi nite time horizon and inventory constraints change the dynamics of collusion compared to previous work. Reward-punishment schemes cannot extend past the end of the episode, making collusion theoretically impossible (with a backward induction argument), but practically achievable (with imperfect learning agents and long enough episodes). In Section 5, we demonstrate efficient computation of the competitive Nash Equilibrium, a challenging task on its own. We show that two common deep RL algorithms, Proximal Policy Optimization (PPO) [46] and Deep Q-Networks (DQN) [39], learn to collude in our model in two distinct ways that align with the intuition provided in Section 4. We analyze the learned strategies,fi nding that agents collude while being aware of the competitive best response, and maintain collusion with a reward-punishment scheme. We show that collusion is robust to changes in agent hyperparameters, unless learning targets are made intentionally unstable, in which case agents converge to a competitive best response strategy. In Section 6, we conclude and discuss future research directions.

2 RELATED WORK

Our work is related to a line of research into competitive and collusive dynamics that emerge between reinforcement learning algorithmic pricing agents in economic games. We refer to Abada et al. [1] for an excellent survey on this topic, and to Appendix C for a more detailed literature review.

Recent research most relevant to us focuses on the Bertrand oligopoly, where agents compete by setting prices and using Q-learning. The main line of research uses Bertrand competition with an infinite time horizon [20], with follow-up work using DQN [32], varying the demand model [6], modeling sequential rather than simultaneous agent decisions [34], or an episodic setting with contexts [24]. Findings reveal frequent, though not universal, collusion emergence, often explained by environmental *non-stationarity* preventing theoretical convergence guarantees. Agents consistently learn to charge supra-competitive prices, punishing deviating agents through 'price wars' before reverting to collusion. The robustness of collusion emergence to factors like agent number, market power asymmetry, and demand model changes underscores the potential risks posed by AI in pricing.

Which factors support and impede the emergence of learned collusion remain debated. Some [2, 54] argue collusion results from agents 'locking in' on supra-competitive prices early on due to insufficiently exploring the strategy space, suggesting a dependence on the choice of hyperparameters. Most studies identifying collusion used Q-learning, with others showing competitive behavior, raising questions about algorithm specificity [45]. However, recent work [21, 36] using PPO in ridesharing markets and infinite Bertrand competition respectively, suggests otherwise. We expand on thesefi ndings in a more realistic episodic,fi nite horizon market with inventory constraints using deep RL algorithms (PPO and DQN), to manage our model's larger state spaces and dynamic environments.

3 PROBLEM STATEMENT

We introduce a multi-agent market model for inventory-constrained goods with a sell-by date, such as perishable items, hotel rooms, or tickets, using airline revenue management (ARM) as an example. We show how to model such markets as a Markov game and define a collusion metric based on the profits achieved under perfect competition and collusion.

3.1 Episodic Markov Games

An *episodic Markov game* [38] is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, T)$ where $\mathcal{S}$ represents the common state space shared by all agents, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the joint action space for n agents, $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the stochastic state transition function, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the reward received by agent i, and T specifies the episode length in discrete timesteps.

At each time step t, agents observe the current state $s_t \in \mathcal{S}$ and simultaneously choose actions following their respective time-dependent policies $\pi_{i,t} : \mathcal{S} \to \mathcal{P}(\mathcal{A}_i)$. We use π_i to denote agent i's vector of policies over time. Each agent's goal is to maximize its cumulative reward over the episode given the game's dynamics,

$$\max_{\pi_i} \sum_{t=1}^{T} R_i(s_t, a_t) \tag{1a}$$

$$\textbf{s.t.} \quad s_{t+1} \sim P(s_t, a_t); \quad a_{j,t} \sim \pi_{j,t}(s_t). \tag{1b}$$

The main challenge infi nding optimal policies in a Markov game is that agent i's optimization problem depends on the actions chosen by all other agents. In a learning context, where agents optimise their policies simultaneously, this optimisation becomes non-stationary and convergence is not guaranteed. For a detailed discussion on the challenges of multi-agent reinforcement learning we refer the reader to a number of surveys [18, 30, 57, 58].

3.2 Markets as Episodic Markov Games

We extend the *Bertrand competition* [11] model, where agents compete to sell a common good. In its simplified one-shot setting, sellers choose prices, and consumers react, deciding which quantity to buy from each seller based on some demand function of those prices. In contrast, we model markets where goods can be sold in multiple timesteps $t = 1, \ldots, T$ over afi nite episode. The Markov game's action space $\mathcal{A}$ consists of the prices agents can set, and an agent's policy π_i represents their pricing strategy. Each timestep t, agents observe the state s_t and simultaneously use their policy π_i to choose an action in the form of a *price* $p_{i,t} = \pi_i(s_t)$, forming the price vector $p_t = (p_{1,t}, \ldots, p_{n,t})$. In the following, we use $p_{i,t}$ for actions instead of $a_{i,t}$ to emphasize that the actions represent prices.

Additionally, we assume that each agent has afinite *capacity* $I_i \in \mathbb{N}$ of goods that they can sell throughout the episode. At each time t, each agent has a remaining *inventory* of tickets $x_{i,t} \in \{0, \ldots, I_i\}$, resulting in an inventory vector $x_t = (x_{1,t}, \ldots, x_{n,t})$. We define the state of the game at time t as the most recent price vector and current inventory, i.e., $s_t = (p_{t-1}, x_t)$. We motivate this definition of the state by the fact that in the non-episodic setting, most recent prices provide agents sufficient information to learn various strategies including perfect competition and collusion [20, 24]. However, investigating the effect of longer recall is an interesting direction for future research.

With prices chosen, a state transition from time t to $t+1$ occurs: For each agent i, the market determines a *demand* $d_{i,t}$, the agent sells a corresponding *quantity* $q_{i,t} = \min(d_{i,t}, x_{i,t})$ bounded by their inventory, and their inventory is updated to $x_{i,t+1} = x_{i,t} - q_{i,t}$. With our choice of demand function (cf. Section 3.4), this transition to the next period's state $s_{t+1} = (p_t, x_{t+1})$ is deterministic. Finally, each agent receives their profit as a *reward* $R_{i,t} := R_i(s_t, p_t) = (p_{i,t} - c_i)q_{i,t}$, with c_i their constant *marginal cost* per good sold.

3.3 Application to Airline Revenue Management

To motivate the episodic Markov game framework, we consider the Airline Revenue Management (ARM) problem. In ARM, agents represent airlines competing to sell a fixed number of seats on a direct flight (also called a *single-leg* flight) between two cities on the same day. The problem is naturally episodic; episodes start when the flight schedule is announced and end at departure, i.e., the sell-by date of the tickets. Furthermore, each airline is constrained by the capacity of their respective aircraft. We consider each route on each day to form a single independent market. Expanding our model to connecting (*multi-leg*) flights, several flights on the same day, cancellations, and overbooking promises interesting future work. This market is a great example with fierce competition, a history of tacit collusion [14], real-time public information on offered ticket prices and inventories via Global Distribution Systems (GDS), and early adoption of dynamic pricing algorithms [35][2].

3.4 Demand Model

We employ a modified *multinomial logit (MNL)* demand model, commonly used in Bertrand price competition [20, 21, 24], to simulate the probability of a customer choosing each agent's product, ensuring demand distribution among all agents rather than clustering on the best offering. The normalized *demand* for agent i's good in period t is

$$d_{i,t} = \frac{\exp\big((\alpha_i - p_{i,t})/\mu\big)}{\sum_{j \in N_t^a} \exp\big((\alpha_j - p_{j,t})/\mu\big) + \exp(\alpha_0/\mu)} \in (0,1), \quad (2)$$

where $N_t^a := \{j \in N \mid x_{j,t} > 0\}$, α_i is agent i's good's quality, α_0 is the quality of an outside good for vertical differentiation, and μ is the horizontal differentiation scaling parameter. The quantity demanded from agent i at time t is then defined as $q_{i,t} = \min\{\lfloor \lambda d_{i,t} \rceil, x_{i,t}\}$, scaling demand with a factor $\lambda \in \mathbb{N}$ and rounding to the nearest integer to account for the sale of goods in whole numbers. We incorporate *choice substitution*, or *demand adaptation*, by summing only over agents with available inventory N_t^a. If an agent is sold out, demand shifts to those with remaining inventory, preventing the sold-out agent's actions from affecting the demand and rewards of others.

[2]Adoption of dynamic pricing algorithms in this industry has historically been limited to low-cost carriers, due to established carriers heavily depending on legacy systems and data-driven forecasting models. See lit. review in Appendix C.

3.5 Measuring Collusion and Competition

We measure the collusion of an observed episode and agent strategies on a scale from 0 (*competitive*) to 1 (*collusive*). First, we establish the two extremes in the Markov game as the competitive Nash equilibrium and the monopolistic optimum that we can later use as reference points for collusion.

Definition 3.1 (Competitive & collusive solutions). A collection of agent policies $(\pi_1, \ldots, \pi_n)$ is called

- *Competitive*, or *Nash equilibrium*, if no agent i can improve their expected episode profit $\mathbb{E}_\pi[\Sigma_{t=1}^T R_{i,t}]$ by unilaterally picking a different policy given fixed opponent policies.
- *Collusive*, or *monopolistic optimum*, if it maximizes expected collective profits, $\mathbb{E}_\pi[\Sigma_{i=1}^n \Sigma_{t=1}^T R_{i,t}]$.

As we argue theoretically in Section 4 and show experimentally in Section 5.1, both admit solutions that feature constant prices across an episode, which we call p^N and p^M for the Nash and monopoly cases, respectively. In our model, the collusive prices p^M are higher than the competitive prices p^N, and the same holds for the correspondingly achieved profits R^M and R^N. At the Nash equilibrium, both unilaterally increasing or decreasing one's price reduces profits. However, if all agents jointly increase prices, the increase in margin outpaces the decrease in (MNL) demand, leading to increased profits for everyone. Building on these two solutions, we define a measure for collusion.

Definition 3.2 (Collusion measure). We define agent i's *episodic profit gain* as

$$\Delta_{i,e} := \frac{1}{T} \sum_{t=1}^{T} \frac{\bar{R}_{i,t} - R_{i,t}^N}{R_{i,t}^M - R_{i,t}^N}.$$

The *episodic collusion index* is measured as the generalized mean of the individual episodic profit gains, i.e.,

$$\Delta_e := \left(\frac{1}{n} \sum_{i=1}^{n} \Delta_{i,e}^\gamma \right)^{\frac{1}{\gamma}}$$

indicating a competitive or collusive outcome at 0 or 1, respectively.

The generalized mean interpolates the arithmetic mean (i.e., average) and geometric mean, which are obtained by setting $\gamma = 1$ and $\gamma = 0$ respectively. We use $\gamma = 0.5$ for our collusion index. Our reason is that the geometric mean has an advantage against the simple average used in previous studies [20, 24], as it more strongly penalizes unilateral competitive defections in a collusive arrangement. However, it interprets any outcome where at least one agent achieves only competitive, or even sub-competitive profits (defining the measure via clamping negative profit gains to zero) as fully competitive, even if others prices above the competitive level and achieve considerable supra-competitive profits. The generalized mean provides a good middle ground. To better interpret negative values, we replace $\Delta_{i,e}^\gamma$ with $\text{sgn}(\Delta_{i,e})|\Delta_{i,e}|^\gamma$. See Appendix E.1 for a comparison of means. Ultimately, how to aggregate the individual profit gains is a subjective question with trade-offs that depend on which outcomes one wants to differentiate the best. E.g., the following outcomes $(\Delta_{1,e}, \Delta_{2,e})$ of $(0.1, 0.1)$, $(0, 0.2)$ or $(-0.1, 0.3)$ have the same average episodic profit gain, but quite different agent behavior and implications on consumer welfare, especially if agents' qualities, costs, and thus equilibrium profits, are not symmetric. Exploring alternative measures, which could be inspired by social choice theory, is a promising avenue for future research.

4 THE COLLUSIVE STRATEGY LANDSCAPE

In this section, we discuss how our model's episodic nature and finite inventory significantly affect the strategies for establishing and maintaining learned tacit collusion compared to the previously considered infinite horizon setting. It is common economic intuition (e.g., [31]) that in order to maintain collusive agreements, agents need to remember past actions and have mechanisms to punish those who deviate from the agreed-upon strategy[3]. Standard punishment strategies include a temporary or permanent shift to a competitive price level after the deviation is detected which results in lower profits for allfi rms. It has been well documented that learning algorithms converge to these strategies in the infinite horizon setting [20, 21, 32]. Such strategies are only credible as long as sufficient time and supply is available for the punishment to offset the short-term gains from a deviation. These conditions are not always met in our settings that lead to new collusive strategies.

Infinite horizon games. These settings allow for deriving unique competitive and collusive equilibrium price levels through implicit formulas with the most commonly used Bertrand competition models. They provide the most room for collusive strategies to emerge and sustain since there is no time constraint for a punishment strategy's credibility. Typically, stable collusion manifests in two forms. First, *reward-punishment schemes:* Agents cooperate by default and punish deviations. A deviating agent is punished by others charging competitive prices, thereby removing the benefits of collusion temporarily, until the supra-competitive prices are reinstated. This dynamic involves agents synchronizing over rounds to restore higher price levels after a deviation. This pattern can be observed as fixed, supra-competitive prices and verified by forcing one agent to deviate and recording everyone else's responses. Second, *Edgeworth price cycles:* This pattern involves agents sequentially undercutting each other's prices until one reverts to the collusive price, prompting others to follow, restarting the undercutting cycle [34].

Episodic games. In comparison to the infinite horizon setting, collusive strategies can now emerge in two distinct ways. First, through *intra-episode* action-based communication, where agents gradually raise their prices through signaling within a single episode. Second, through training *across many episodes*, where agents eventually learn policies that implement collusive pricing immediately from the start of each new episode. The latter form is prevalent in oligopolistic settings and possibly explained by learners overfitting their strategies to familiar opponents. When faced with new opponents, collusive agents initially play competitively before reestablishing collusion through continued learning [24]. This robustness result suggests thatfi rms aiming to collude can pretrain their pricing agents separately, needing only (likely legal) alignment on the high-level training setups (e.g., algorithm classes, observation modeling, exploration schedule). In our experiments in Section 5.5, we observe evidence of both types of collusion.

Thefi nite time horizon restricts collusive potential by limiting the efficacy of reward-punishment schemes used in infinite-horizon games to maintain collusion. In a one-shot game ($T = 1$) in our Bertrand setting, there exists a unique Nash equilibrium at the competitive price level, as unilateral deviation from collusive prices is profitable and future punishment is impossible. In thefi nite horizon case ($T > 1$), the same logic applies at thefi nal period ($t = T$), such that any Nash equilibrium strategy will price competitively in the last timestep. By induction from $t = T$ backwards, this argument extends to all periods $t = T - 1, \ldots, 1$, defining a unique Nash equilibrium where agents compete throughout the episode. Does this mean that collusion in episodic games is impossible? No: If agents remember past interactions across episodes, deviations can be punished in future episodes. Surprisingly, our experiments in Section 5 show that even without cross-episode memory, learning agents in sufficiently long episodes can converge to collusive strategies of the signaling, stable or cyclic kind. We observe that some agents learn to play collusively at episode start and defect toward the end, suggesting that discovering the full backward induction argument through (often random) exploration is unlikely enough in practice.

Episodic, inventory-constrained model. Inventory constraints significantly complicate the state and strategy space by making the reward achieved from a pricing strategy dependent on inventory levels. Determining the competitive and collusive price levels becomes more complex because the solution formulas from the Bertrand or Cournot settings require smoothness or convexity assumptions that no longer hold, preventing the standard uniqueness proofs. We approachfi nding a Nash equilibrium by modeling each episode as a simultaneous-move game where agents set entire price vectors before the episode starts for the complete episode. We provide further details in Section 5.1. We solve the resulting generalized Nash equilibrium problem numerically and prove that its solutions are Nash equilibria in our Markov game. Wefi nd that in our model, both the competitive and collusive solutions consist of repeating their prices from the one-period equivalents T times. If agents discount future rewards, both equilibria shift to lower prices and higher profits early in the episode and vice versa toward its end. In addition, price levels remain distinct even with strict inventory constraints. Due to the difficulty in predicting or interpreting observed behavior in this complex setting, we see value in analyzing different types of learners as part of future work.

5 EXPERIMENTS

In Section 5.1 wefi rst show how tofi nd the competitive and monopolistic price levels needed to calculate the collusion measure defined in Definition 3.2, and how they change under different inventory constraints. Then, we show that PPO [46] and DQN [39], two commonly used deep RL algorithms, can learn to collude in our episodic model. Finally, we analyze their learned strategies and their dependence on hyperparameters.

[3]Recent work [5] suggests that there can exist stable, collusive equilibria of strategies that do not encode threats. They show that near-monopoly prices can arise if afi rst-moving agent deploys a no-regret learning algorithm, and the second agent subsequently picks a non-responsive pricing policy.

5.1 Obtaining Competitive and Collusive Equilibrium Prices

Previous works' Bertrand settings use analytic formulae to compute Nash equilibrium and monopolistic optimum price vectors p^N and p^M for single-period cases. However, a closed-form solution is not available for our problem setting. We therefore use numerical methods to calculate the competitive and collusive solutions as defined in Definition 3.1 and use these values to define the collusion measure in Definition 3.2.

First, we calculate the profits and prices in the monopolistic (perfectly collusive) setting by assuming a central optimizer who chooses prices for all agents maximizing the total profit. Second, to calculate the same for the competitive Nash equilibrium, we model an entire episode as a *simultaneous-move game (SMG)*, where all agents i must simultaneously decide all T prices in their vector $p_i = (p_{i,1}, \ldots, p_{i,T})$ before an episode begins. Let $p = (p_1, \ldots, p_n)$ encompass all agents' price vectors, with p_{-i} representing all agents' vectors except i's. The solution to this SMG is then a Generalized Nash Equilibrium defined as follows.

Definition 5.1. The *Generalized Nash Equilibrium Problem (GNEP)* consists offi nding the price vector $p^* = (p_1^*, \ldots, p_n^*)$ such that for each agent i, given p_{-i}^*, the vector p_i^* solves the following inventory-constrained revenue maximization problem

$$\begin{aligned} \max_{p^{(i)}} \quad & \sum_{t=1}^{T} (p_{i,t} - c_i) \lfloor \lambda d_{i,t} \rfloor \\ \text{subject to} \quad & \sum_{t=1}^{T} \lfloor \lambda d_{i,t} \rfloor \leq I, p_i \geq 0. \end{aligned} \tag{3}$$

The solution price vector p^* can be interpreted as the *actions* of a set of agent policies playing an episode of the Markov game. The following lemma shows that a set of policies that result in the price vector p^* form a Nash Equilibrium in the Markov Game.

LEMMA 5.2. *Given a Markov Game with deterministic transitions, let $p^* = (p_1^*, \ldots, p_n^*)$ be the solution to Equation (3) and define $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$, as $\pi_i^*(s_t) = p_{i,t}^*$ for all i, t, and $s_t \in \mathcal{S}$. Then π^* is a Nash equilibrium in the Markov Game.*

The full proof can be found in Appendix D. Details of our numerical approach to solving the GNEP are found in Appendix A.

Without discounting, the episodic equilibrium price vectors repeat the single-period equilibrium with the same parameters T times. Figure 1 shows how inventory constraints affect market dynamics. When inventories exceed the demand at the competitive equilibrium, the equilibria correspond to the unconstrained setting. As inventories shrink, the competitive price level rises, as it is harder forfi rms to undercut and profit from the increased demand. When inventory size matches the demand at the collusive price, the collusive and competitive price levels converge. Further tightening of constraints pushes both coinciding prices higher. In our experiments we choose the constraint's value between the two extremes to allow for differentiation between competitive and collusive behavior and a well-defined collusion index, and investigate the effect of the inventory size on learned collusion in Section 5.6.

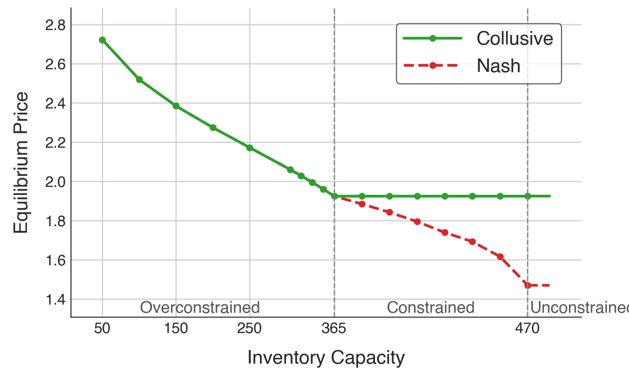

Figure 1: One-period equilibrium price levels as a function of inventory capacity for two equally constrained agents.

5.2 Model Parameters

We evaluate the potential for RL algorithms to collude in our model using a duopoly situation with two agents[4]. We use either of two popular algorithms, namely Deep Q-Networks (DQN) [39] and Proximal Policy Optimization (PPO) [46] for learning without weight-sharing between agents. Agents represent identicalfi rms, sharing the same qualities $\alpha_i = 2$, marginal costs $c = c_i = 1\ \forall i$, a horizontal differentiation factor of $\mu = 0.25$, an outside good quality of $\alpha_0 = 0$, and a demand scaling factor of $\lambda = 1000$. For the main results presented in Section 5.4 and Section 5.5, we set the inventory constraints to $440 \cdot T$ and the episode length $T = 20$.

Due to the symmetry between agents, Nash and monopolistic price levels are identical for both of them, and the price levels and the corresponding demands are $p^N = 1.693, p^M = 1.925$ and $d^N = 440 d^M = 365$ for our inventory constrained case. Agents choose prices from a discretized interval $[p^N - \xi(p^M - p^N), p^M + \xi(p^M - p^N)]$ with 15 steps and $\xi = 0.2$, such that the competitive and collusive actions correspond to $a^N = 2$ and $a^M = 12$ respectively. In particular, the price range for our setting is $[1.693, 1.925]$. In Appendix E.3, we provide further results on experiments with a price range defined with the unconstrained Nash equilibrium prices to demonstrate that agents are still capable of learning collusion and their actions quickly converge to the price range defined with the constrained Nash equilibrium prices.

5.3 Training Setup

We train our algorithms by playing 1000 and 50,000 episodes for PPO and DQN, respectively, and updating weights after every episode for PPO or every fourth for DQN. We train 100 pairs of PPO or DQN on unique random seeds (40 for the boxplots). After training, we analyze each agent pair by observing their play in a single episode. This joint training aligns with previous work [20, 36] and real market situations, wherefi rms learn while competing, updating pricing strategies based on market success. Solid lines and shaded

[4]Two agents suffice to demonstrate learned collusion in thefi nite horizon game and the impact of inventory constraints. A duopoly is a reasonable assumption in the ARM domain, as many routes are dominated by 2-3 airlines. For $n > 2$ agents, [2, 32] show collusion indeed diminishes due to exponential growth in joint policy space hindering joint exploration, but does not fully disappear.

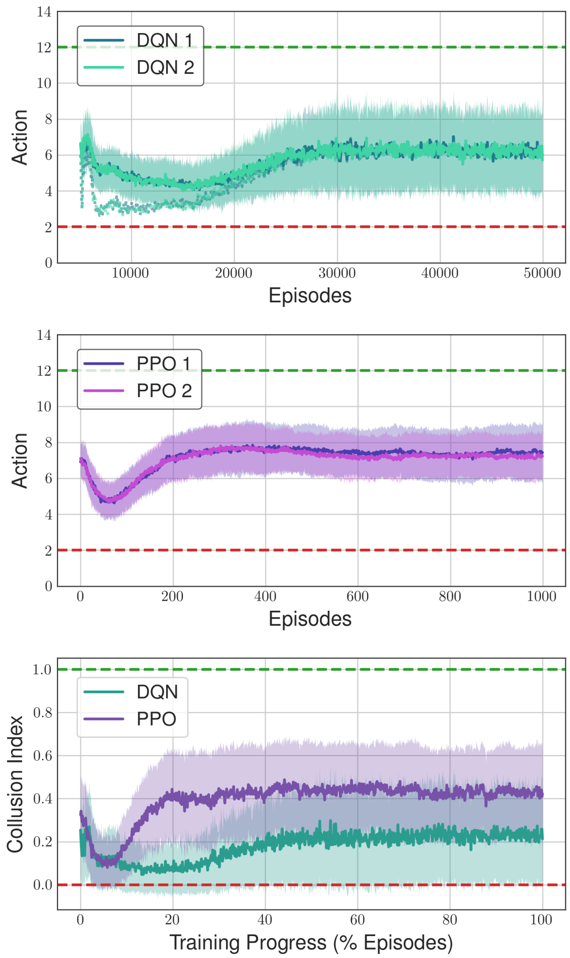

Figure 2: Evolution of training two DQN and two PPO agents in our model, showing average agent actions per episode (top, middle) and collusion index (bottom) with collusive (green) and competitive (red) actions indicated. The dotted lines are DQN's greedy actions. Both DQN and PPOfi rst converge to competition before gradually rising toward collusion.

areas in our plots represent the averages and standard deviations of their metrics. For our DQN agent, we use *epsilon-greedy* exploration with an exponentially decaying epsilon, while the PPO agent anneals its entropy coefficient to similarly reduce exploration over time. For evaluation episodes, DQN uses a fully greedy action selection. We normalize the rewards during training to the interval $[0, 1]$ based on minimum and maximum possible values. This makes training slightly more stable. However, collusion is still achieved with unnormalized rewards. A full description of the hyperparameters used for DQN and PPO can be found in Appendix B.2. We use the JAX framework on a custom codebase built on [56]. Our experiments were run on a compute cluster on a mix of nodes with each run using at most four vCPU cores, 8GB of RAM, and either a NVIDIA T4 or NVIDIA V100 GPU. However, a single run can be done on a consumer laptop (Apple M1 Max, 32GB RAM) in under one hour.

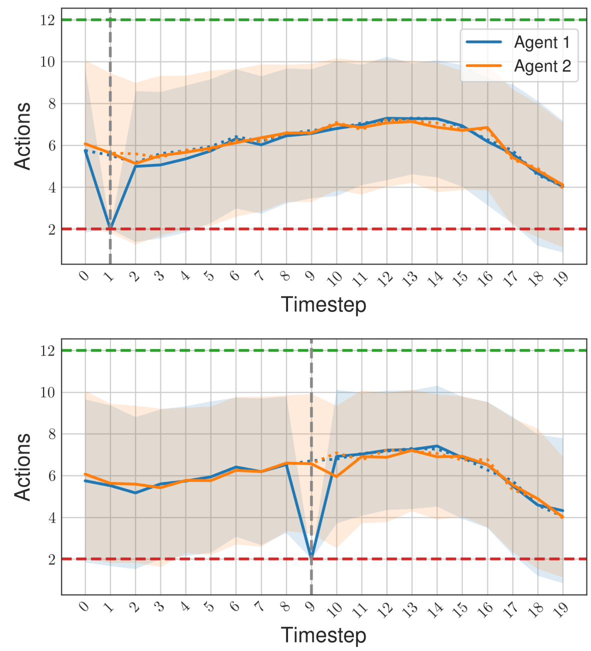

Figure 3: Behavior of two DQN agents during an episode after forcing one agent to deviate at time $t = 1$ and $t = 9$ respectively. Dotted lines indicate evolution without deviation. Deviations provoke a competitive reaction, with both agents quickly returning to collusion.

5.4 Analysis of Learning Process

Figure 2 shows two training runs for DQN and PPO agents. For both algorithms, agents quickly converge to each other and to competition as their learning targets are initially unstable, with high epsilon (DQN) and entropy (PPO) forcing random actions. This makes it hard for agents to adapt to their opponent's underlying policy and leads to them learning the best-response strategy against a random opponent, playing competitively. As the exponentially decaying epsilon and entropy curvesfl atten and the agents face an increasingly predictable opponent that they can adapt to, they begin colluding. Prices rise gradually and jointly before leveling off at a collusive level. PPO converges in both much fewer episodes and achieves higher levels of collusion, with an average collusion index of $\Delta_e = 0.43$ over the last 10% of episodes, compared to DQN's $\Delta_e = 0.23$. These values, lower than in prior studies in the standard Bertrand setting [20, 21, 32], highlight the greater challenge of collusion in our more complex model. Regulatory efforts could focus on the gradual increase in prices to mitigate algorithmic collusion, which we consider to be an interesting direction for future work.

5.5 Analysis of Collusive Strategies

After training, we simulate the agents in an evaluation episode (Figure 3). We focus on DQN here, discussing PPO in Appendix E.2. Our DQN agents show behavior that slowly rises in collusiveness until both agents defect near the end of the episode. This suggests that the agents are capable of learning that late defection cannot be

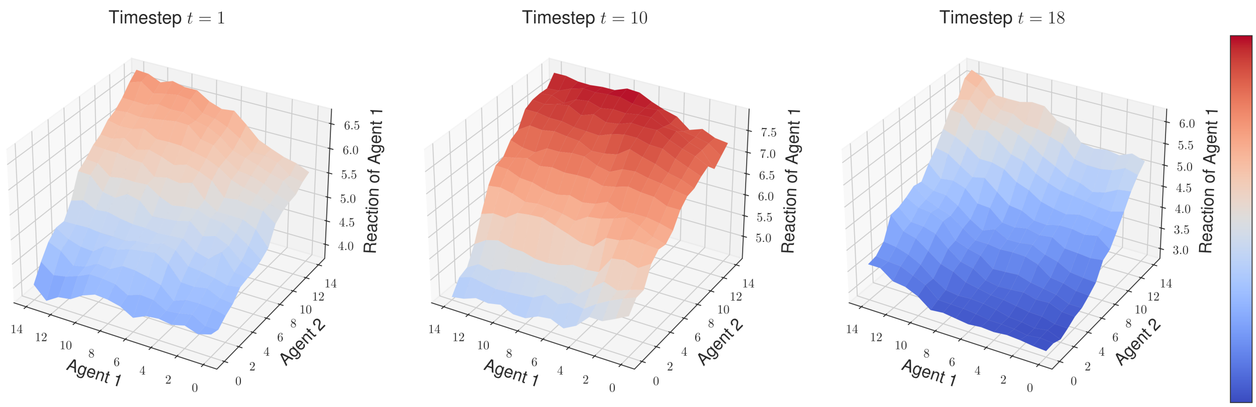

Figure 4: The surfaces show a DQN agent 1's learned best response under their greedy policy (i.e., the action with the highest Q-value) to a state given by both agent's prices (x- and y-axes), timestep and symmetric remaining inventory level.

punished, while not fully applying the backward induction argument from Section 4. The rise in collusion at the beginning of the episode suggests a capability of establishing *intra-episode* collusion, with the gradual, mutual price increase acting as a form of signaling. In Appendix E.8, we show results without inventory constraints, where the agents' price curve isfl atter, suggesting a strategy more based on solidified collusion over multiple episodes.

To analyze the nature of the learned strategy, we force one agent to deviate at a certain timestep and record the response by both agents similarly to Calvano et al. [20]. Interestingly, deviation produces only a small reaction by the competing agent, while the deviating agent quickly returns to near their collusive level. With a deviation at time $t = 1$ or at $t = 9$, the impact on overall episode profits is negligible for both agents, with the deviating agent breaking even and the non-deviating agent losing only 0.2% profit overall. We refer the reader to Appendix E.9 for details.

Figure 4 shows the best-response surfaces of thefi rst agent at different points in the episode, with a remaining inventory linearly interpolated from full to none over the episode (corresponding to the agents' evaluated strategy) and averaged over 100 trained agent pairs. The agent always punishes opponent deviations by pricing lower than the previous price level. At episode start and end, the agent's best-response surface shows some symmetry indicative of more competitive behavior, reacting to their own deviations with even lower prices, anticipating a 'price war'. During the middle, they instead return to previous or even higher collusion levels after own defections, signaling cooperation, and punish opponent deviations with slight undercutting. Near the end, they shift to more competitive behavior, punishing deviations more strongly. This topology suggests that if both agents start near the competitive equilibrium, they will both react in a way that jointly 'climbs the hill' to collusion, leveling out at an action of roughly 7 as indicated by theflat top. The second agent behaves similarly. These results suggest that DQN agents are well aware of competitive strategies and choose to collude in a robust way reliant on rewards and punishments. Appendices E.4 and E.5 contain results for uneven inventory constraints and limiting observability of opponent inventory and time, neither of which significantly hinder the emergence of collusion.

5.6 Hyper- and Environment Parameters

We analyze the impact of changing agent hyperparameters and environment characteristics on the convergence and collusive tendencies of DQN and PPO agents. We show comparisons for agent learning rate, inventory constraint, and episode length here, with additional results deferred to Appendix E.10. To judge the convergence of two agents toward each other throughout the training run, we use the following metric:

$$\frac{1}{0.1E} \sum_{e=0.9E}^{E} \frac{1}{T} \sum_{t=1}^{T} \frac{|p_{0,t} - p_{1,t}|}{p^M - p^N} \tag{4}$$

adapted from Deng et al. [21], where E is the number of training episodes. It takes the average difference of both agents' prices across an episode relative to the width of the Nash-monopolistic price interval. Values below 0.2 are interpreted as converged.

In our analysis, we vary single parameters from the reference setup described in Section 5.2, train agents on 40 different seeds, and for each parameter value, record the distribution of convergence metric and collusion index over those seeds, averaged over the last 10% of training run episodes.

Learning rate is perhaps the most important agent parameter, as it regulates the impact of all other agent parameters. Figures 5a and 5b demonstrate that both PPO and DQN agents achieve better convergence and increased tendency to compete at lower learning rates. The reduced ability to adapt to an opponent's strategy still allows agents to learn the opponent-independent best-response of competition at initial training episodes, but attempts to establish the gradual, mutual increase in price seen in Figure 2 happen more rarely and revert to competition more often. A higher learning rate does not translate to more likely collusion, as the increased ability to adapt to an opponent is balanced by the potential to overreact to the opponent's random actions. Overall, collusion and convergence appear to be robust to moderate changes in learning rate.

We compare metrics among different initial inventory sizes in Figure 5c. Inventory sizes shown are per-timestep; a value of 440 represents a total inventory size of $440 \cdot T$, which we use for the other results. Smaller inventories show better convergence and more competitive behavior for both PPO and DQN. This has geometric intuition (cf. Appendix E.6): visualize each agent's reward landscape as a surface over the grid of both agents' prices. Each agent tries to climb toward their peak on the side of the grid's diagonal where they undercut their opponent. Steps toward their peak along their axis harm their opponent. To achieve collusion, agents must jointly climb the ridge along the diagonal of the grid where their landscapes intersect. The closer the two agent's peaks are to the monopolistic optimum on the diagonal, the smaller their incentive to deviate and the smaller the negative impact on their opponent from deviation, easing cooperation. Decreasing inventory capacities reduces the range of prices that agents are incentivized to use as the Nash equilibrium price approaches the monopolistic price. In this "zoomed in" part of the price grid, the peaks now appear further away from each other, making the coordination problem harder.

Figure 5d shows the effect of changing episode lengths. As conjectured in Section 4, longer episodes increase collusion tendencies for both types of learners by providing more opportunities to punish deviations. While PPO's convergence is unaffected, DQN's convergence suffers. This is expected, as DQN generally scales worse to larger state spaces than PPO. It relies on accurately estimating the expected reward for each state-action pair and sufficiently exploring the state space, which becomes harder as that space grows.

We identified additional hyperparameters affecting collusion, such as PPO's number of training epochs (higher increases collusion) and DQN's buffer size (larger increases collusion), shown in Appendix E.10. It is possible to hinder collusion by introducing instability in learning targets, e.g., by filling DQN's buffer or PPO's rollouts with experiences gathered from 'parallel environments'. This parallelization is commonly done to increase training speed on accelerator hardware, but has a concrete impact in this model. We demonstrate this with PPO in Appendix E.7.

6 CONCLUSION

We formulate price competition between producers as an episodic Markov game motivated by Airline Revenue Management (ARM) and facilitating the analysis of tacit collusion within a finite time horizon and inventory-constrained markets. We propose numerical methods to find competitive and collusive solutions in our model due to the lack of analytical solutions and define a collusion metric based on the total profit achieved in a full episode. Our analysis shows that collusion consistently emerges between independent DQN and PPO algorithms after a brief period of competition and that trained agents quickly revert back to collusive prices after a forced deviation. The proven collusive potential of RL agents in our setting covering many real markets reinforces the call for the development of mitigation strategies and regulatory efforts [19].

We see our work as a first step toward understanding pricing competition in markets like airline tickets, hotels, and perishable goods with future research directions in extending our Markov Game model to domain specifics. Additionally, we see a need to consider multi-agent specific algorithms, e.g., opponent-shaping agents [50], that could establish stronger collusion or even exploit market participants, significantly harming social welfare.

ACKNOWLEDGMENTS

B. Pásztor is supported by an ETH AI Center doctoral fellowship.

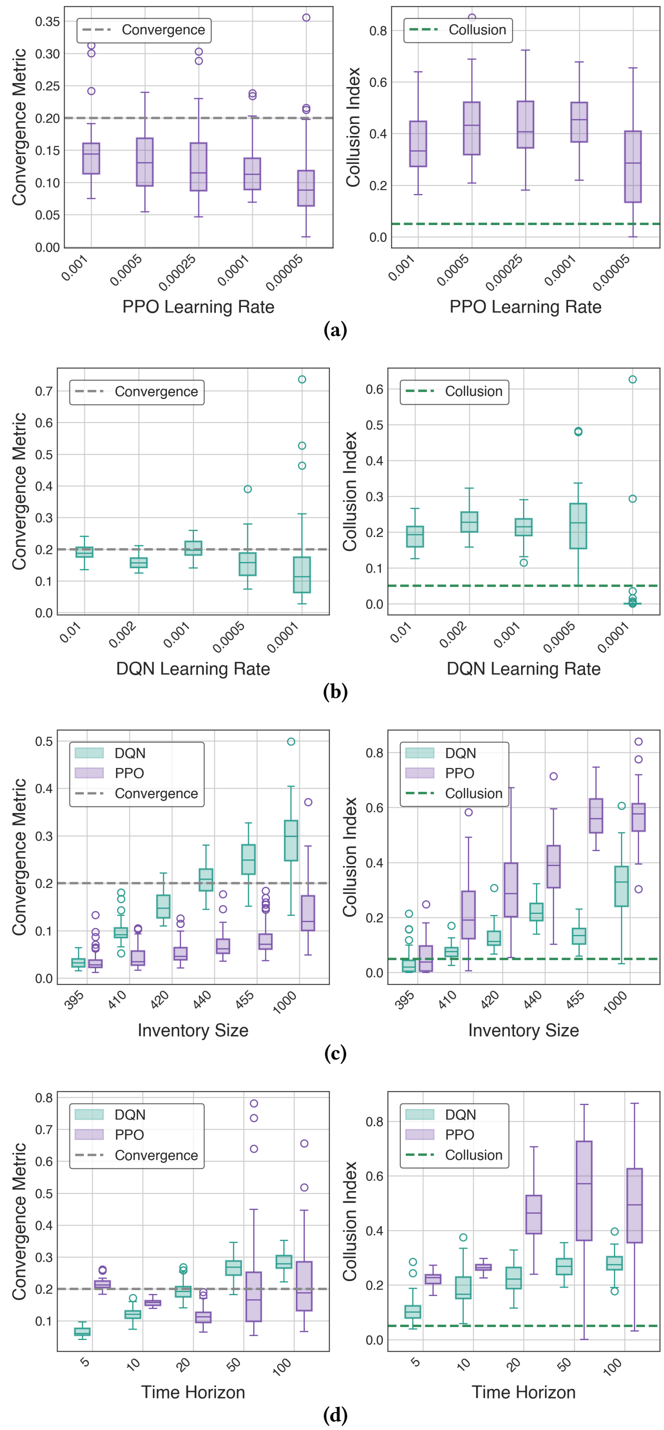

Figure 5: Convergence and collusion metrics for DQN and PPO training runs with varied learning rate (a) & (b), inventory sizes (c), and episode time horizons (d). Initial inventory size is the value shown, times the time horizon T. Collusion is robust against varying (yet sufficiently large) learning rate. Longer episodes show less reliable convergence, higher potential collusion due to more effective punishment strategies.

REFERENCES

[1] Ibrahim Abada, Joseph E. Harrington Jr, Xavier Lambin, and Janusz M. Meylahn. 2024. Algorithmic Collusion: Where Are We and Where Should We Be Going?

[2] Ibrahim Abada and Xavier Lambin. 2023. Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided? *Management Science* 69, 9 (2023), 5042–5065.

[3] Rodrigo Acuna-Agost, Eoin Thomas, and Alix Lhéritier. 2023. Price Elasticity Estimation for Deep Learning-Based Choice Models:An Application to Air Itinerary Choices. In *Artificial Intelligence and Machine Learning in the Travel Industry: Simplifying Complex Decision Making*, Ben Vinod (Ed.). Springer Nature Switzerland.

[4] Neda Etebari Alamdari and Gilles Savard. 2021. Deep Reinforcement Learning in Seat Inventory Control Problem: An Action Generation Approach. *Journal of Revenue and Pricing Management* 20, 5 (2021), 566–579.

[5] Eshwar Ram Arunachaleswaran, Natalie Collina, Sampath Kannan, Aaron Roth, and Juba Ziani. 2024. Algorithmic Collusion Without Threats. arXiv:2409.03956

[6] John Asker, Chaim Fershtman, and Ariel Pakes. 2022. Artificial Intelligence, Algorithm Design, and Pricing. *AEA Papers and Proceedings* 112 (2022), 452–56.

[7] Stephanie Assad, Robert Clark, Daniel Ershov, and Lei Xu. 2024. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Journal of Political Economy* 132, 3 (2024), 723–771.

[8] Lawrence M Ausubel. 1991. The failure of competition in the credit card market. *The American Economic Review* (1991), 50–81.

[9] Peter Paul Belobaba. 1987. Air Travel Demand and Airline Seat Inventory Management. *Flight Transportation Laboratory Reports* (1987).

[10] Francisco Beneke and Mark-Oliver Mackenrodt. 2021. Remedies for Algorithmic Tacit Collusion. *Journal of Antitrust Enforcement* 9, 1 (2021), 152–176.

[11] Joseph Louis François Bertrand. 1883. Review of "Theorie mathematique de la richesse sociale" and of "Recherches sur les principles mathematiques de la theorie des richesses.".

[12] Dimitris Bertsimas and Sanne de Boer. 2005. Simulation-Based Booking Limits for Airline Revenue Management. *Operations Research* 53, 1 (2005), 90–106.

[13] Nicolas Bondoux, Anh Quan Nguyen, Thomas Fiig, and Rodrigo Acuna-Agost. 2020. Reinforcement Learning Applied to Airline Revenue Management. *Journal of Revenue and Pricing Management* 19, 5 (2020), 332–348.

[14] Severin Borenstein and Nancy L Rose. 1994. Competition and price dispersion in the US airline industry. *Journal of Political Economy* 102, 4 (1994), 653–683.

[15] Gianluca Brero, Eric Mibuari, Nicolas Lepore, and David C. Parkes. 2022. Learning to Mitigate AI Collusion on Economic Platforms. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 35. 37892–37904.

[16] Juan José Miranda Bront, Isabel Méndez-Díaz, and Gustavo Vulcano. 2009. A Column Generation Algorithm for Choice-Based Network Revenue Management. *Operations Research* 57, 3 (2009), 769–784.

[17] Bundeskartellamt and Autorité de la Concurrence. 2019. *Algorithms and Competition*. Technical Report.

[18] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1* (2010), 183–221.

[19] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, Joseph E. Harrington, and Sergio Pastorello. 2020. Protecting Consumers from Collusive Prices Due to AI. *Science* 370, 6520 (2020).

[20] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110, 10 (2020), 3267–3297.

[21] Shidi Deng, Maximilian Schiffer, and Martin Bichler. 2024. Algorithmic Collusion in Dynamic Pricing with Deep Reinforcement Learning. arXiv:2406.02437

[22] Joris Dinneweth, Abderrahmane Boubezoul, René Mandiau, and Stéphane Espié. 2022. Multi-Agent Reinforcement Learning for Autonomous Vehicles: A Survey. *Autonomous Intelligent Systems* 2, 1 (2022), 27.

[23] Directorate-General for Competition (European Commission), Yves-Alexandre de Montjoye, Heike Schweitzer, and Jacques Crémer. 2019. *Competition Policy for the Digital Era*. Publications Office of the European Union.

[24] Nicolas Eschenbaum, Filip Mellgren, and Philipp Zahn. 2022. Robust Algorithmic Collusion. arXiv:2201.00345

[25] European Union. 2012. Treaty on the Functioning of the European Union. Arts. 101-109.
http://data.europa.eu/eli/treaty/tfeu_2012/oj.

[26] European Union. 2019. Regulation (EU) 2019/712 on safeguarding competition in air transport, and repealing Regulation (EC) No 868/2004.
http://data.europa.eu/eli/reg/2019/712/oj.

[27] Francisco Facchinei and Christian Kanzow. 2007. Generalized Nash Equilibrium Problems. *4OR* 5, 3 (2007), 173–210.

[28] David Genesove and Wallace P Mullin. 2001. Rules, communication, and collusion: Narrative evidence from the sugar institute case. *American Economic Review* 91, 3 (2001), 379–398.

[29] Abhijit Gosavi, Naveen Bandla, and Tapas K. Das. 2002. A Reinforcement Learning Approach to a Single Leg Airline Revenue Management Problem with Multiple Fare Classes and Overbooking. *IIE Transactions* 34, 9 (2002), 729–742.

[30] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review* 55, 2 (2022), 895–943.

[31] Joseph E. Harrington. 2018. Developing Competition Law for Collusion by Autonomous Artificial Agents. *Journal of Competition Law & Economics* 14, 3 (2018), 331–363.

[32] Matthias Hettich. 2021. Algorithmic Collusion: Insights from Deep Learning. *SSRN Electronic Journal* (2021).

[33] Alexander Kastius and Rainer Schlosser. 2022. Dynamic Pricing under Competition Using Reinforcement Learning. *Journal of Revenue and Pricing Management* 21, 1 (2022), 50–63.

[34] Timo Klein. 2021. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics* 52, 3 (2021), 538–558.

[35] Oded Koenigsberg, Eitan Muller, and Naufel Vilcassim. 2004. *EasyJet Airlines: Small, Lean and with Prices that Increase over Time.* Working Paper. London Business School Centre for Marketing. https://lbsresearch.london.edu/id/eprint/3369/

[36] Pravesh Koirala and Forrest Laine. 2024. Algorithmic Collusion in a Two-Sided Market: A Rideshare Example. arXiv:2405.02835

[37] Ryan J. Lawhead and Abhijit Gosavi. 2019. A Bounded Actor–Critic Reinforcement Learning Algorithm Applied to Airline Revenue Management. *Engineering Applications of Artificial Intelligence* 82 (2019), 252–262.

[38] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning (ICML'94)*. 157–163.

[39] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533.

[40] Leon Musolff. 2022. Algorithmic pricing facilitates tacit collusion: Evidence from e-commerce. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. 32–33.

[41] Maureen K. Ohlhausen. 2017. *Should We Fear The Things That Go Beep In the Night? Some Initial Thoughts on the Intersection of Antitrust Law and Algorithmic Pricing.* Technical Report. Federal Trade Commission.

[42] Rupal Rana and Fernando S. Oliveira. 2014. Real-Time Dynamic Pricing in a Non-Stationary Environment Using Model-Free Reinforcement Learning. *Omega* 47 (2014), 116–126.

[43] Rupal Rana and Fernando S. Oliveira. 2015. Dynamic Pricing Policies for Interdependent Perishable Products or Services Using Reinforcement Learning. *Expert Systems with Applications* 42, 1 (2015), 426–436.

[44] Pouria Razzaghi, Amin Tabrizian, Wei Guo, Shulu Chen, Abenezer Taye, Ellis Thompson, Alexis Bregeon, Ali Baheri, and Peng Wei. 2022. A Survey on Reinforcement Learning in Aviation Applications. arXiv:2211.02147

[45] J. Manuel Sanchez-Cartas and Evangelos Katsamakas. 2022. Artificial Intelligence, Algorithmic Competition and Market Structures. *IEEE Access* 10 (2022), 10575–10584.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347

[47] Syed A.M. Shihab and Peng Wei. 2022. A Deep Reinforcement Learning Approach to Seat Inventory Control for Airline Revenue Management. *Journal of Revenue and Pricing Management* 21, 2 (2022), 183–199.

[48] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[49] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the Game of Go without Human Knowledge. *Nature* 550, 7676 (2017), 354–359.

[50] Alexandra Souly, Timon Willi, Akbir Khan, Robert Kirk, Chris Lu, Edward Grefenstette, and Tim Rocktäschel. 2023. Leading the Pack: N-player Opponent Shaping. arXiv:2312.12564

[51] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction.* MIT press.

[52] Kalyan T. Talluri and Garrett J. Van Ryzin. 2004. *The Theory and Practice of Revenue Management.* International Series in Operations Research & Management Science, Vol. 68. Springer US.

[53] Garrett van Ryzin and Jeff McGill. 2000. Revenue Management Without Forecasting or Optimization: An Adaptive Algorithm for Determining Airline Seat Protection Levels. *Management Science* 46, 6 (2000), 760–775.

[54] Ludo Waltman and Uzay Kaymak. 2008. Q-Learning Agents in a Cournot Oligopoly Model. *Journal of Economic Dynamics and Control* 32, 10 (2008), 3275–3293.

[55] Rui Wang, Xianghua Gan, Qing Li, and Xiao Yan. 2021. Solving a Joint Pricing and Inventory Control Problem for Perishables via Deep Reinforcement Learning. *Complexity* (2021).
[56] Timon Willi, Akbir Khan, Newton Kwan, Mikayel Samvelyan, Chris Lu, and Jakob Foerster. 2023. Pax: Scalable Opponent Shaping in JAX. https://github.com/ucl-dark/pax.
[57] Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. 2023. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review* 56, 6 (2023), 5023–5056.
[58] Yaodong Yang and Jun Wang. 2021. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. arXiv:2011.00583