Lange, Fabian; Schlosser, Rainer

Article

Dynamic pricing with waiting and price-anticipating customers

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: Lange, Fabian; Schlosser, Rainer (2025) : Dynamic pricing with waiting and price-anticipating customers, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 14, pp. 1-20, https://doi.org/10.1016/j.orp.2025.100337

This Version is available at:
https://hdl.handle.net/10419/325814

Operations Research Perspectives 14 (2025) 100337

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Dynamic pricing with waiting and price-anticipating customers

Fabian Lange[1], Rainer Schlosser *

Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

ARTICLE INFO

Dataset link: https://anonymous.4open.science/r/StrategicCustomerRL-6C1E

Keywords:
Strategic customers
Dynamic pricing
Reference prices
Price anticipation
Reinforcement learning

ABSTRACT

Over the last decades, dynamic pricing has become increasingly popular. To solve pricing problems, however, is particularly challenging if the customers' and competitors' behavior are both strategic and unknown. Reinforcement Learning (RL) methods are promising for solving such dynamic problems with incomplete knowledge. RL algorithms have shown to outperform rule-based competitor heuristics if the underlying Markov decision process is kept simple and customers are myopic. However, the myopic assumption is becoming increasingly unrealistic since technology like price trackers allows customers to act more strategically. To counteract unknown strategic behavior is difficult as pricing policies and consumers buying patterns influence each other and hence, approaches to iteratively update both sides sequentially are time consuming and convergence is unclear. In this work, we show how to use RL algorithms to optimize prices in the presence of different types of strategic customers that may wait and time their buying decisions. We consider strategic customers that (i) compare current prices against past prices and that (ii) anticipate future price developments. To avoid frequently updating pricing policies and consumer price forecasts, we endogenize the impact of current price decisions on the associated changes in forecast-based consumer behaviors. Besides monopoly markets, we further investigate how the interaction with strategic consumers is affected by additional competing vendors in duopoly markets and present managerial insights for all market setups and customer types.

1. Introduction

1.1. Dynamic pricing and strategic customer behavior

Enterprises can utilize dynamic pricing strategies by adjusting prices frequently to changes in demand, resource prices, and other relevant factors. Digital sales environments significantly increase the sales data available to incumbent companies [1]. Insights into customer behavior, particularly responses to different offer prices, may be exploited by a seller adopting offering prices to maximize profits [2]. Dynamic pricing strategies emerged in various industries, including airline ticketing, electricity prices, gasoline pricing, or hotel fees and are studied extensively, see, *e.g.*, [3–5], and [6].

Facing dynamic prices, customers have the opportunity to study past offers – as reference prices, cf. [7], or in order to anticipate future prices – and postpone their purchase, strategically aiming for the lowest offer price [8]. Strategic customer behavior is studied as part of investigations on dynamic pricing [2] and self-contained [9,10] but leaves a variety of open research questions as pointed out in the survey by [11].

Enabling the simulation of pricing strategies and their outcomes, dynamic programming (DP) or RL techniques can be leveraged to calculate or predict a pricing policy that leads to the highest possible profit [2]. Therefore, the complexity of the actual market is reduced to some elementary features that are modeled suitably. Based on the outcomes of the simulation in an environment created as a digital twin of a real market, to some extent, conclusions from the optimal policy can be drawn to be later applied in the real world.

In this work, we aim to extend studies on strategic customer behavior and its influence on optimized dynamic pricing strategies. For this reason, types of consumer strategies for postponing a purchase to achieve a higher consumer rent are investigated. Our work targets to optimize pricing strategies under different customer behaviors defined by reasonable rules, *i.e.*, buying policies building on varying availability of information on past prices, and study the results for vendors and customers. Furthermore, we seek to explore how such pricing strategies should be further adapted in extended scenarios with additional duopoly competition.

In this context, we study whether RL methods can be used to optimize dynamic pricing strategies in such scenarios. As a customer

* Corresponding author.
E-mail addresses: fabian.lange@student.hpi.de (F. Lange), rainer.schlosser@hpi.de (R. Schlosser).
[1] Authors contributed equally.

https://doi.org/10.1016/j.orp.2025.100337
Received 3 October 2024; Received in revised form 3 March 2025; Accepted 26 March 2025
Available online 11 April 2025

base, we examine four different types of customers with backward and forward-looking purchase behavior. We seek to model the problem such that RL-based solutions converge after a sufficient number of training steps and find approximated equilibria between vendor and customer actions.

Finally, the ability to compute optimized pricing policies and to analyze the associated interplay with consumers' and competitors' strategies is beneficial for researchers and practitioners. The results can help industry vendors to better understand strategic interaction in specific markets and to use inferred insights to maximize profits by counteracting strategic customer behavior in practice. On the other hand, consumers also benefit from our analysis as best-performing buying strategies can be identified by our evaluations as well.

1.2. Contributions

The optimization of pricing strategies in the presence of strategic consumers under incomplete information and additional competitive merchants is a challenging and understudied problem.

While small Markov Decision Process (MDP) problems with full information can be solved by standard DP methods, already slightly more complex ones can only be solved approximately. If – as in practical applications – complete information about the model dynamics is not available, the problem gets further difficult, and effective learning techniques are required. In this regard, advanced RL techniques can be a suitable alternative as they are able to address more complex MDPs with incomplete information. However, latest developments in computer science, primarily RL, have rarely been tested in dynamic pricing problems in the presence of strategic customers and competition.

Particularly, optimized pricing policies for scenarios with an unknown share of customers – that actively *time* their sales while dynamically comparing prices to older reference prices or to anticipated future prices – have hardly been studied and it remains unclear to which results such optimized policies lead to for both merchants and consumers.

In this context, our paper studies how to determine optimized pricing policies for competitive markets with patient consumers even though unknown to the seller. Our contributions can be summarized as follows.

- We study the mutual interplay of a self-learning dynamic pricing agent, a competing merchant, and unknown mixtures of myopic, price-aware, and price-anticipating consumers. We show how to formulate the problem within an MDP market environment such that RL techniques can be successfully applied in various market scenarios where the agent has no initial knowledge of the underlying environment.
- Our model allows to optimize counter-strategies against different types of buying strategies. These buying strategies can include the following strategic components: (i) the ability to wait for better prices, (ii) comparing current prices to older reference prices, and (iii) price anticipations.
- While existing work is mostly limited to one or two-period models when modeling reference prices and price predictions, we are able to optimize prices in the presence of more realistic consumer behaviors that take longer sequences of periods into account. This allows us to study consumers that are influenced by longer price patterns and that are able to anticipate future price developments for multiple periods ahead.
- Although changes in the agent's policy and consumers demand influence each other, we avoid mutual subsequent updates of the agent's beliefs in demand on the one hand and the consumers' beliefs in future prices on the other. Instead, we use a synchronized learning process. At the agent's side, we endogenize the ability to learn the impact of single price decisions on changes in consumer demand. On the customer side, we use an auto-regressive price prediction approach that works with comparably few data. This allows controlling the impact of price decisions on both expected future profits as well as associated changes in consumers' demand behavior via their influence on price forecasts.
- Further, we are able to extend the monopoly setup to a duopoly market against another competing merchant. Thereby, we can study the influence of a competitor on both the optimized pricing policy of the agent and the customers purchase timing.
- We provide extensive numerical evaluations for different types of customers in monopoly as well as duopoly setups. Besides recurring and reference price-based customer behaviors, we particularly study how to counteract mixtures of price-anticipating and myopic customers.
  Finally, we are able to analyze and compare how the shares of certain types of non-myopic consumers affects the consumer rent of myopic consumers as well as the rewards of both competing merchants. Moreover, we can compare different backward and forward-looking customer behaviors regarding the performance at the consumer side.
- We provide an open-source simulation and evaluation framework, see code repository https://anonymous.4open.science/r/StrategicCustomerRL-6C1E/.

The remainder of this paper is organized as follows. In Section 2, we discuss related works. In Section 3, we introduce our market model and propose the modeling of different types of forward- and backward-looking customer behavior as well as a competitive vendor, which strategically responds to the agent's prices. Further, we describe how to embed RL methods in the proposed MDP environment. In Section 4, we present our evaluation results for various scenarios, compare the performances of all market participants, and infer managerial insights. In Section 5, we summarize the main results obtained, discuss limitations of our models, and provide ideas for future research. Section 6 concludes the paper.

2. Related work

This section provides a research overview on dynamic pricing (Section 2.1) and strategic customer behavior (Section 2.2). We present related works as an intersection of findings from operations research and management science, economics, and computer science. In Section 2.3, we describe the research gap our paper seeks to address.

2.1. Dynamic pricing

Dynamic pricing can be defined as identifying optimal prices or a pricing strategy for unchanged products such as goods or services for different points in time or customer groups, according to [12,13].

The topic of dynamic pricing has received much attention in recent years. Originating in the first description of optimal price calculations in economic settings by [14], it is applied to various use cases and studied by different scientific communities today. [2] concludes that the operations research and management science cluster generally aims to find a seller's optimal pricing policy. At the same time, economists try to explain price formation and buying behavior in markets. The computer science community was found to study more complex market simulations that are no longer tractable with mathematical analyses. At this point, complex machine learning techniques are conducted to find optimal policies.

[15] provide a recent literature review on competitive online retail and categorize whether a product competes with identical or differentiated products, the influence of time dependency on the simulation of a market, and the overall market structure. They found that substitute competition is high because of the higher number of product offers but did not neglect the possible effect of differentiated products. Furthermore, the authors identified that time-dependent pricing policies

outperform their static counterparts. Regarding the market structure, Gerpott and Berends discovered that most studies use a monopolistic or duopolistic setup for simplification of the in-reality oligopoly structured market.

Early works, see, *e.g.*, [16–18] for a monopoly setup or [19,20] for duopoly markets, often assume stylized demand dynamics, a full information setup, or myopic consumer behavior to be able to identify optimal pricing policies analytically.

In contrast, the growing influence of the computer science community increased interest in more complex models that cannot be solved analytically anymore. Instead, approaches based on machine learning were put forward to enable the inclusion of more influencing factors towards a more realistic market implementation.

In this context, [21] concluded that RL allows for solutions to previously intractable problems. They highlight the appropriate fit between the nature of RL algorithms optimizing a reward in the long run and a vendor optimizing revenue in a market.

The studies of [22,23] used RL algorithms to simulate a market environment with multiple vendors and complex dynamics. Further, there are more applied simulations in the areas of airline ticketing [24], inventory management [25], electricity prices [26], or regenerative electric heating [27]. Most recent applications of deep RL in dynamic pricing were made by [28] pricing access to express lanes, [29] using Proximal Policy Optimization (PPO) in a ride pricing problem, [30] using a Soft Actor Critic (SAC) algorithm for electric vehicle charging prices, or [31] applying Deep Deterministic Policy Gradient (DDPG) as well in the vehicle charging domain.

2.2. Strategic customer behavior

Rational consumer behavior and particularly, the subarea of strategic customer behavior used in dynamic pricing models is widely studied. Strategic customer behavior can be defined as anticipating future price changes and adjusting purchase timing accordingly [32]. Originating in 1972, [33] discovered that even a monopoly vendor has to offer its product at the margin cost without achieving any profit when facing strategic customers.

[10] present classifications for studies on strategic customer behavior. The *capacity* indicates the amount of products a vendor can sell. It can be limited or infinite and additionally used for decision-making. Prices can be determined at the beginning or during the selling horizon, *i.e.*, *time of pricing*. The pricing policy can be classified into markup, markdown, or a combination. A markup policy can only increase prices to react to market change, while markdown policies are limited to decreasing prices. The majority of papers use a combination of both. A *demand arrival process* can be simultaneous at the beginning of the selling horizon or sequential during the horizon. The *number of time periods* can be discrete or continuous and finite or infinite. Most publications act in a finite discrete time setting. Customers and vendors can consider the time in their decision-making or not, *i.e.*, *time preference*. The *market setup* can be a monopoly, duopoly, or oligopoly. Furthermore, the criterion of *information setting* indicates the degree of available information to customers and vendors. In their summary, they mention the issue of comparing the results of different studies.

[32] put forward a dynamic pricing model with intertemporal demand, having customers maximize their utility by choosing whether to accept the current price offer of a monopolistic vendor, postpone their buying to wait for a better offer, or exit the market. Contrary to intuition, they found that the ability to wait enables the vendor to benefit.

The more recent survey by [11] categorizes research on strategic customer behavior into three mechanisms (Pricing, Inventory, and Information) and summarize existing strategies that have been developed to mitigate the profit decrease of a company with strategic customers for all of them. Furthermore, they conclude that strategic customers' behavior as well as vendors' policies to counteract have to be researched via learning methods with regard to increasing amounts of available sales data.

Many existing works often use analytical tractable models (*e.g.* [34]), which as an advantage allow for analytic results, but on the downside limiting market complexity to a certain degree and requiring vendors to have crucial information on demand and market dynamics. Additionally, in earlier investigations, considered customer behaviors inhibit only a small degree of strategy by considering only one or two last periods (*e.g.* [35]) influencing the buying policy, which, in turn, does not allow to model consumers' price anticipation in a fully realistic manner.

2.3. RL-based learning approaches and research gap

[11] query to research how strategic customers could leverage information on past prices to monitor prices or predict future price trajectories. Despite the high availability of studies on strategic customer behavior in dynamic pricing research, there is comparably few related work on applying learning techniques, like state-of-the-art RL methods. While multiple usages of RL in dynamic pricing are studied solely, see Section 2.1, the combination with strategic customer behavior is scarce to our knowledge.

[36] use an aggregating algorithm to dynamically price a perishable product and simultaneously learn the market in a monopoly setup. [37] study a Stackelberg game under a two-period model to learn a single consumers valuations of a product using Bayesian learning techniques.

One of the first approaches using Deep Q-networks (DQN) and SARSA to learn pricing policies and ordering quantities in the presence of strategic consumers (in a monopoly market) was done by [38], including the modeling of substitute products. Their calculation with markdown prices for used products in the second of two combined decision periods and the inclusion of inventory considerations show that DQN and SARSA are suitable for solving large-scale pricing optimizations. However, the customer's strategy employed in this study can be explained as the choice between buying a new or a used product, without explicit anticipating of future prices or consideration of past prices. The works by [39,40] use deep RL techniques to study joint dynamic pricing and inventory control problems with reference price effects. Price anticipations of customers or competing sellers are not considered.

Overall, how to counteract strategic customer behavior with forward and backward-looking variants has not sufficiently been studied, probably since classical DP methods are not suitable given the problem's complexity and information structure.

RL methods seem to be a promising alternative but have been comparably rarely been used in this domain as they cannot be applied in a straightforward way. To be able to successfully apply RL methods to markets with complex strategic consumers their behavior and observable state components have to be modeled with care.

In this paper, we show how to model different strategic consumer behaviors, including reference price-based as well as price-anticipating ones, in a way such that state-of-the-art RL methods can be applied. Most importantly, we avoid to use data-intensive price forecasts on the consumer side as the sequential interplay of mutually updating pricing policies at the vendor's side and the consumer's price forecast is time consuming and its convergence remains unclear. Instead, we endogenize the consumer's price forecast based on current price histories within a tractable state space. This way, no iterative adaption of pricing policies and price forecasts is necessary as the impact of current price decisions on rewards and particularly also the associated change in the forecast-based consumer behavior can be taken into account.

3. Model description

In this section, we define our market model as an MDP (Section 3.1) and model four different types of strategic customers (Section 3.2). In Section 3.3, we describe how to embed RL methods in the proposed MDP environment and identify suitable algorithms to be used in the evaluation.

3.1. Market environment

In the following, we describe the market setup, a firm's admissible controls, competitors' reactions, the consumer arrival process, and a firm's objective. A code repository is available.[2]

3.1.1. Setup

We consider an infinite time horizon with discrete time periods. We consider $n_{vendors}$ competing firms. Each firm sells a (standardized) product. We simplify our market to one offered product without inventory constraints such as storage cost or a limited number of available products.

3.1.2. A firm's controls and competitors' reactions

Each firm k, $k = 1, \dots, n_{vendors}$, sets a price $p^{(k)} \in A$, for his/her product, where A denotes the set of admissible prices with a maximum offer price a_{max}. We mainly take the perspective of firm 1 and seek to apply a self-learning strategy, while – in case of competition – other competitors use rule-based strategies. The set of admissible price can be modeled continuously or discrete, providing the agent vendor with a set of possible actions.

Taking the perspective of one specific firm, *e.g.*, firm $k = 1$, this firm sets its price at the beginning of a period of length one, *e.g.*, from time t to $t + 1$, taking into account the current prices $\vec{p}(t) \in A^{n_{vendors}}$ of all competing firms at time t.

Within the period $(t, t + 1)$ each firm $k = 2, \dots, n_{vendors}$ of the competing $n_{vendors} - 1$ firms adjusts its price at its corresponding point in time $\tau^{(k)} \in (t, t+1)$ in a similar way by reacting to the current prices at time $\tau^{(k)}$, *i.e.*, $\vec{p}(\tau^{(k)})$.

In this context, for each firm, we assume non-anticipating Markovian strategies, where all competitors' current prices are observable.

As price response strategies, in general, deterministic as well as probability distributions over admissible prices can be used. Further, besides current prices also, *e.g.*, the historic prices of the last h periods up to time t (denoted by $H_t^{(h)}$) could be used as input. Competitors' strategies are mutually not observable by the merchants.

3.1.3. Modeling of consumer behavior

We consider a stream of arriving consumers whose number, type, and timing can be defined in a steady deterministic or in a random fashion. We consider n_c different non-exclusive types of customers. Further, we allow for seasonal demand with cycle length $n_{seasons}$, *i.e.*, for all consumer types demand in period t is characterized by the season (*e.g.* the weekday, month, or the season of the year), where we use the cyclic formulation, $t = 0, 1, \dots$,

$$i_{season}(t) = t \bmod n_{seasons} \in \{0, 1, \dots, n_{seasons} - 1\}. \tag{1}$$

A single consumer of type c, $c = 1, \dots, n_c$, arriving at a certain point in time t observes the current offer prices $\vec{p}_t := (p^{(1)}(t), \dots, p^{(n_{vendors})}(t))$ for all $n_{vendors}$ firms.

The choice behavior of consumers of type c can be defined arbitrarily and may include a no-buy option (cf. $k = 0$ as no firm make a sale). In our model, we assume that a single customer buys at most one product and that the buying behavior of consumer type c in period t is expressed as a probability distribution for buying no item at all (cf. $P_t^{(c,0)}(\vec{p}_t; H_t^{(h)}) \geq 0$) or buying a product from firm k (cf. $P_t^{(c,k)}(\vec{p}_t; H_t^{(h)}) \geq 0$), $k = 1, \dots, n_{vendors}$, given the current prices $\vec{p}_t \in A^{n_{vendors}}$ (and a price history $H_t^{(h)}$) such that for all $c = 1, \dots, n_c$ and $t = 0, \dots, n_{seasons} - 1$, we have

$$\sum_{k=0,1,\dots,n_{vendors}} P_t^{(c,k)}(\vec{p}_t; H_t^{(h)}) = 1. \tag{2}$$

Note, while for myopic consumers demand only depends on the current prices $\vec{p}_t$, for certain consumer types c we allow that their buying probabilities at time t, cf. (2), additionally depend on last historic prices, cf. $H_t^{(h)}$, up to a certain age h (which may vary with c).

3.1.4. Problem formulation from a single firm's perspective

A firm k's rewards, $k = 1, \dots, n_{vendors}$, are characterized by its sales. By $i_t^{(k)}$, we denote the number of items sold to firm k within period $(t, t + 1)$, which are obtained from the number of realized sales $Y_t^{(c,k)}$ of firm k for customers of type c, cf. (2).

We allow that a firm's policy may depend on current competitors' prices as well as a recent history $H_t^{(n_{last})}$ involving the last n_{last} periods. Given a pricing policy $p_t^{(k)} = a_t^{(k)}(\vec{p}_t; H_t^{(n_{last})})$, a firm k's random accumulated future profits from time t on (discounted on time t) amount to, $t \geq 0$, $k = 1, \dots, n_{vendors}$,

$$G_t^{(k)} := \sum_{j=t}^{\infty} \gamma^{j-t} \cdot i_j^{(k)} \cdot p_j^{(k)}, \tag{3}$$

where the discount factor for one time period is γ.

Our firm 1's goal is to determine a non-anticipating (Markovian) feedback pricing policy that for a given initial state s_0 characterized by current market prices $\vec{p}_0$ and historic prices $H_0^{(n_{last})}$, *i.e.*,

$$s_0 := \left(\vec{p}_0, H_0^{(n_{last})}\right), \tag{4}$$

maximizes the expected total discounted rewards, cf. (3), from time $t = 0$ on:

$$E\left(G_0^{(1)} \mid s_0\right). \tag{5}$$

Due to the size of the state space, standard dynamic programming (DP)-based solution techniques (also assuming complete information about the dynamics of the underlying process) are, in general, not applicable.

Hence, we seek to apply RL algorithms to the problem as an alternative approach. Note, this is possible as long as states, actions, rewards, and state transitions of the MDP can be expressed in a standardized way - the so-called environment. Before embedding and selecting suitable RL algorithms, see Section 3.3, we first introduce the modeling of different classes of strategic behaviors within our MDP model.

3.2. Description of different types of customer behaviors

In Section 3.2.1–3.2.4, we describe the overall modeling of four different versions (types) of customer's behavior in detail. Therefore, we define specific strategies characterized by demand probabilities $P_t^{(c,k)}(\vec{p}_t; H_t^{(h)})$, cf. (2), in the presence of a given price history H_t and a certain waiting behavior. Note, this formulation allows for using arbitrary choice models.

Further, in our model, all customer types except the myopic one are able to postpone their buy and to wait for better prices. Next, we define the customer behaviors based on the demand function of the myopic customer and add different specific rules to decide whether to wait or not. Note, consumer types are not exclusive and some of their components can overlap.

3.2.1. Myopic customer

In general, we consider a "myopic customer", cf. type $c = 0$, with seasonal demand characterized by probabilities $P_t^{(0,k)}(\vec{p}_t)$, cf. (2). In case of no sale, the consumer leaves the market and does not wait for better prices or systematically checks prices in future periods.

[2] https://anonymous.4open.science/r/StrategicCustomerRL-6C1E/

3.2.2. Recurring customer

We model a so-called "recurring customer", cf. type $c = 1$, exactly as the myopic customer, cf. $P_t^{(1,k)}(\vec{p}_t)$, with the extension of being able to postpone his/her buy. For this reason, when simulating/evaluating the model, we consider a waiting pool for types of recurring consumers. All customers who are drawn not to accept one of the current price offers enter this pool and may return the next timestep. Note, the size of the waiting pool is not observable for the vendors.

3.2.3. Price-aware customer

A so-called "price-aware consumer", cf. type $c = 2$, acts reference price based and evaluates current prices against historic prices. Hence, within the probabilities $P_t^{(2,k)}(\vec{p}_t; H_t^{(\delta)})$, we can formulate decision rules such as "*buy if and only if the best current price within $\vec{p}$ is at least x% lower than the lowest price within the prices of the last δ periods, cf. $H_t^{(\delta)}$, and the current offer is below a certain upper threshold price*". As this class of behavior includes waiting options, we also consider a corresponding waiting pool.

3.2.4. Anticipating customer

For modeling the so-called "anticipating customer", cf. type $c = 3$, we leverage a basic auto-regressive (AR) approach to predict future minimum prices. Here, we assume that customers know the number of seasons (cf. weekly or monthly cycles). Note, the cycle length of a seasonal effect could also be discovered automatically (*e.g.*, by an autocorrelation analysis in case of more complicated seasonal patterns).

In each timestep t, we consider the set of current prices $\vec{p}_t$ and the latest price history $H_t^{(\phi)} = (\vec{p}_{t-1}, \dots, \vec{p}_{t-\phi})$ to fit an auto-regressive model $AR(H_t^{(\phi)})$ of order ϕ. As customers look for best prices, in our modeling, we consider the time series of minimum prices over all firms to predict best future prices. Alternatively, the AR model could also be applied for each firm k separately to fit and predict prices on a firm level.

We denote $\varphi_0, \dots, \varphi_\phi$ as parameters and ϵ as error term of $AR(H_t^{(\phi)})$. The parameters can be adjusted using an ordinary least-squares approach to minimize the difference between predicted and actual values. As a result, we obtain a simple model that allows predicting a one-dimensional time series. Considering a firm's last ϕ period's prices p_h until time t as well as current price p_t, *i.e.*, $h = t - \phi, t - \phi + 1, \dots, t$, the price p_{t+1} at $t + 1$ of the next timestep can be estimated by z_{t+1} with

$$z_{t+1}(H_t^{(\phi)}) = AR(H_t^{(\phi)}) = \sum_{h=0}^{\phi} \varphi_{(h \mod n_{seasons})} \cdot p_{t-h}. \tag{6}$$

Furthermore, by using the value z_{t+1}, we can also predict the price p_{t+2} of the second-next timestep via z_{t+2}, and so on. This enables us to predict prices of even more prospective timesteps. We model the price-anticipating customer to predict at time t the next θ period's prices.

Based on such predictions (for each firm or minimum prices over time) this type of customer may follow decision rules such as "*buy if there is no price prediction for the next θ periods that is better than the current offer and the current offer is below a certain upper threshold price*". To account for uncertainty or risk-averse effects future price predictions can be easily combined with certain mark-ups or penalties.

Finally, such decision rules can be formulated within probabilities $P_t^{(3,k)}(\vec{p}_t; H_t^{(\phi)})$ as long as the required input for prediction, cf. $\vec{p}_t$ and $H_t^{(\phi)}$, is contained. Due to the waiting options, we also consider a corresponding waiting pool for the price-anticipating consumer type.

3.3. Application of RL methods and agent selection

3.3.1. RL environment formulation

Based on the model description and problem formulation given in Section 3.1, we describe how different RL algorithms can be applied to our problem by mapping the proposed MDP market model to a standard RL framework.

Standard RL frameworks usually require a discrete-time (turn-based) setup and are characterized by a so-called environment, which includes states, actions, reward signals, and state transition dynamics. The RL agent plays against the environment by choosing actions from a certain action space and receiving (aggregated) reward signals and associated state transitions.

From firm 1's, *i.e.*, the agent's perspective, the *observable state* s_t at time t is characterized by the current prices of the competitors as well as the price history (up to a certain number of periods). Note, waiting pools, which are part of the full state ($\tilde{s}_t$) of the environment (to be used for simulation) are – as in practice – not observable to the vendors.

Further, a firm's *action* is simply its current offer price. Hence, for an RL agent, the action space is given by the price set A.

The *reward signal* of a firm is the aggregated reward associated to realized sales (within one period), which is characterized by the underlying customer behavior, cf. (2), including the defined arrival streams of interested consumers of different types, see Section 3.2.

Finally, *state transitions for $\tilde{s}_t$ (including s_t)* are organized via the MDP described in Section 3.1 and governed by the evolution of buying and waiting customers, cf. Section 3.2, as well as the subsequent price adjustments of all competing firms. This, in general, requires that certain, *e.g.*, rule-based, policies are assigned to the competing firms, see Section 3.2.1.

The agent's objective is to find a price update strategy depending on observable states s_t that maximizes expected discounted long-term rewards, cf. (5). Note, the agent does not know internals of the environment, *i.e.*, the defined consumer behaviors, their mixture, and the defined competitors' strategies.

Finally, within the described environment, different standard RL algorithms can be applied by using common RL libraries.

3.3.2. RL algorithm selection

Potential state-of-the-art RL algorithms for our problem are so-called Q-Learning-based techniques and policy gradient algorithms.

Here, we will focus on RL algorithms using neural networks; note that tabular methods (as used in classical dynamic programming) cannot handle the problem because the size of the state space exceeds the computational limits by far.

Further, as also the action space of our problem will be typically large, we decided to consider algorithms that use a continuous action space. The main reason for this is that for a discrete formulation, neural networks require an output neuron for each individual action. However, the size of a discrete action space $|A|$ becomes large with fine-grained price levels.

In line with this, for example, [41] found that Soft Actor Critic (SAC), cf. [42], performed better than Deep-Q-Learning (DQN), cf. [43], on their pricing benchmarks.

Finally, as RL methods with continuous action space, we tested the following state-of-the-art RL algorithms to be applied to our problem: DDPG [44], TD3 [45], A2C [46], SAC, and PPO [47].

For our kinds of experiments PPO performed best and thus, we selected PPO for our evaluation. Further, as we look for solutions avoiding tedious tuning, we use the default hyperparameters and test their suitability for our problem, see Appendix A.1, Table 7.

4. Evaluation

Our evaluation is organized as follows.

In Section 4.1, we define exemplary test strategies for the competitor and, in particular, specify sales probabilities and decision rules for the different types of consumer behaviors. Then, in Section 4.2, we explain the market simulation, define the experimental setup, specify the observable state used by the agent, and provide details regarding the training process and final performance evaluations.

In Experiment 1, as our baseline model, cf. Section 4.3, we study how to optimize prices in the presence of myopic customers with

seasonal demand. In Experiment 2, cf. Section 4.4, we examine the performance of RL agents in case of a share of recurring customers.

In Experiment 3, cf. Section 4.5, we consider price-aware customers (cf. reference price effects). In Experiment 4, cf. Section 4.6, we study how an RL agent learns to counteract in scenarios in which a certain share of the consumer base is price-anticipating.

Note, neither the type of an arriving customer is observable nor the share of non-myopic customers is known to the agent. In all experiments, we consider monopoly as well as duopoly settings against an undercutting competitor. Further, we study examples with time-homogeneous and seasonal demand.

4.1. Example strategies for competitors and consumers

In this section, we describe the competitor's strategy used in duopoly setups (Section 4.1.1), the arrival of consumers (Section 4.1.2), and finally define all 4 types of consumer behaviors to be used in our evaluation (Section 4.1.3–4.1.6). To be able to assess whether self-learning agents are able to obtain reasonable and effective pricing strategies in unknown and competitive markets, we consider synthetic customer behaviors mimicking backward and forward-looking consumers that actively time the purchases.

4.1.1. Competitor's strategies

To challenge the agent with an aggressive opponent, we consider a standard undercutting competitor, which is offering a lower price than the current market price and, therefore, his/her policy depends on the other vendor's offer price. We model the competitor vendor to always undercut the agent vendor's price p by a certain threshold Δ in every timestep but not lower than a certain minimum price a_{floor}. Hence, the response policy denoted by $a^{comp}(p)$, $p \in A$, is:

$$a^{comp}(p) = \begin{cases} p - \Delta & p - \Delta \geq a_{floor} \\ a_{floor} & \text{otherwise.} \end{cases} \tag{7}$$

The competitor's strategy is not known to the agent (and could also be chosen differently).

4.1.2. Consumer arrivals

We assume a constant amount of customers i_{max} arriving at each period. Regarding the multiple types of customers, we determine how many customers arrive for each type as follows. We use probabilities $C(c)$ for an arriving customer to be of type c, $c = 0, 1, \dots, n_c$, *i.e.*,

$$\sum_{c=0,1,\dots,n_c} C(c) = 1, \tag{8}$$

where $c = 0$ refers to myopic customers.

We then randomly draw the absolute number of arriving customers per type from a multinomial distribution based on $C(c)$. Also, the shares $C(c)$ are unknown to the merchants.

4.1.3. Myopic customer behavior

As a test example, we define the buying probabilities of myopic customers (type $c = 0$) motivated by standard multinomial logit (MNL) models. Given a price p, $p \in A$, we calculate a (time-dependent) utility score $u_t(p)$ by, $t = 0, \dots, n_{seasons} - 1$,

$$u_t(p) = \frac{-\alpha \cdot e^{p-\beta_t} - p}{\alpha + \beta_t}, \tag{9}$$

where α and β_t, $t = 0, \dots, n_{seasons} - 1$, are choosable scale parameters. Finally, we use the softmax function to define probability values $P_t^{(0,k)}(\vec{p})$ for any $\vec{p} \in A^{n_{vendors}}$, $t = 0, \dots, n_{seasons} - 1$, $k = 1, \dots, n_{vendors}$,

$$P_t^{(0,k)}(\vec{p}) = \frac{e^{u_t(p^{(k)})}}{\sum_{j=0}^{n_{vendors}} e^{u_t(p^{(j)})}}, \tag{10}$$

where $u_t(p^{(0)}) := u^{(0)}$ corresponds to the case of no sale, cf. $k = 0$. In the duopoly case, we assign arriving myopic customers uniformly distributed to the first and the second half period, *i.e.*, before or after the competitor's price update.

4.1.4. Recurring customer

We model the recurring customer (type $c = 1$) based on the myopic customer, cf. $P_t^{(1,k)}(\vec{p}) := P_t^{(0,k)}(\vec{p})$, see Section 4.1.3, with the extension of being able to postpone his/her buy. With a certain probability π_{remain} all recurring customers who do not to accept one of the current price offers enter/remain in a waiting pool. If a recurring consumer buys a product, he/she leaves the waiting pool. In the next period, besides new arriving consumer of recurring type, every customer in the waiting pool of size $w_t^{(1)}$ returns to the market (with a certain probability π_{return}) to consider a sale again. Hence, on average, a share of $1 - \pi_{return}$ waiting consumers leaves the market (which, *e.g.*, reflects waiting costs or consumer disappointment, cf. [48]. As an upper bound for the waiting pool we use a comparably large value w_{max}.

Note, in the duopoly case, the waiting pool is organized identically. Here, we let new and returning recurring customers arrive uniformly distributed in the first and the second half period, *i.e.*, before or after the competitor's price update.

4.1.5. Price-aware customer

For the price-aware customer (type $c = 2$), we use a decision rule, which is based on a comparison of the current price(s) and the minimum prices of the last δ periods are used as input, cf. $H_t^{(\delta)}$. While in the monopoly case, we simply have $H_t^{(\delta)} = \left(p_{t-1}^{(1)}, p_{t-2}^{(1)}, \dots, p_{t-\delta}^{(1)}\right)$, in the duopoly case the consumers consider the minimum of both competitors' prices for all half-periods within the last δ periods, *i.e.*, $H_t^{(\delta)} = \left(\min_{k=1,2}\left(p_{t-0.5}^{(k)}\right), \min_{k=1,2}\left(p_{t-1}^{(k)}\right), \dots, \min_{k=1,2}\left(p_{t-\delta}^{(k)}\right)\right)$ with $2 \cdot \delta$ entries.

The decision rule for price-aware consumers is characterized by $P_t^{(2,k)}(\vec{p}_t; H_t^{(\delta)}) := \pi_2^{(k)}(\vec{p}_t; H_t^{(\delta)})$, $k = 1, \dots, n_{vendors}$, where

$$\pi_2^{(k)}(\vec{p}_t; H_t^{(\delta)}) = \begin{cases} 1 & p_t^{(k)} \leq \min_{\substack{i=1,\dots,n_{vendors} \\ i=1/n_{vendors},\dots,\delta}} \left\{p_{t-i}^{(k)}\right\} \cdot h_{rel\_ref} \quad \text{and} \quad p_t^{(k)} \leq WTP_{max} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Note, the rule uses the minimum price over the last δ periods as a reference price. If multiple firms k qualify for a sale, we use a uniform draw. Here, the relative threshold $0 \leq h_{rel\_ref} \leq 1$ determines how much cheaper – compared to last prices – current prices have to be, such that a sale is considered; WTP_{max} is a given threshold for the maximum willingness-to-pay (it could also be randomized). In case of no sale, similar to the recurring customers, the price-aware customers use their waiting pool of size $w_t^{(2)}$ as well as probabilities π_{remain} and π_{return}.

Again, in case of a duopoly, we let new and returning price-aware customers arrive uniformly distributed in the first and the second half period and consider a corresponding different price history on a half period level.

4.1.6. Anticipating customer

For the price-anticipating customer (type $c = 3$), we use the AR approach, cf. (6), described in Section 3.2.4, where the minimum prices of the last ϕ periods are used as input, cf. $H_t^{(\phi)}$. While in the monopoly case, we simply have $H_t^{(\phi)} = \left(p_{t-1}^{(1)}, p_{t-2}^{(1)}, \dots, p_{t-\phi}^{(1)}\right)$, in the duopoly case the consumers consider the minimum of both competitors' prices for all half-periods within the last ϕ periods, *i.e.*, $H_t^{(\phi)} = \left(\min_{k=1,2}\left(p_{t-0.5}^{(k)}\right), \min_{k=1,2}\left(p_{t-1}^{(k)}\right), \dots, \min_{k=1,2}\left(p_{t-\phi}^{(k)}\right)\right)$ with $2 \cdot \phi$ entries.

Hence, we use price forecasts for the next θ periods for the minimum price over all firms k in i periods from now, *i.e.*, $z_{t+i}(H_t^{(\phi)})$, where we consider $i = 1/n_{vendors}, 2/n_{vendors}, \dots, \theta - 1/n_{vendors}, \theta$. Based on these forecasts, we let the price-anticipating consumers use the following decision rule, *i.e.*, $P_t^{(3,k)}(\vec{p}_t; H_t^{(\phi)}) := \pi_3^{(k)}(\vec{p}_t; H_t^{(\phi)})$, $k = 1, \dots, n_{vendors}$, where

$$\pi_3^{(k)}(\vec{p}_t; H_t^{(\phi)}) = \begin{cases} 1 & p_t^{(k)} \leq \min_{i=1/n_{vendors},\dots,\theta} z_{t+i}(H_t^{(\phi)}) \quad \text{and} \quad p_t^{(k)} \leq WTP_{max} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

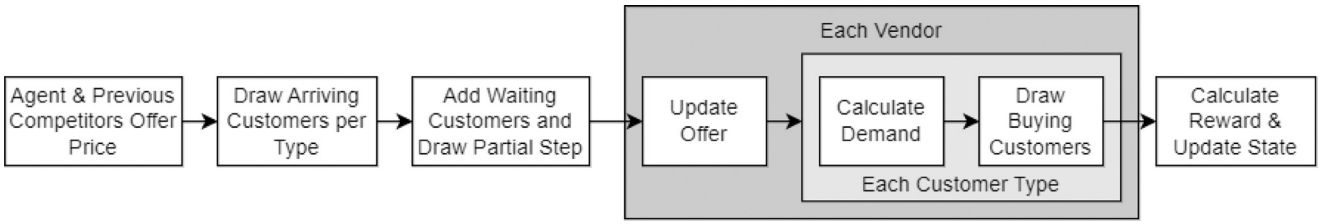

Fig. 1. Overview of the simulation sales and price updates within one period.

If multiple firms k qualify for a sale, we use a uniform draw. Price-anticipating customers also use a waiting pool, cf. size $w_t^{(3)}$, and follow the probabilities π_{remain} and π_{return}. Although the predicted price minimum might be more than one period away, we assume that the consumer may check the next period's market situation to potentially update his/her price forecast.

Further, similarly to the price-aware consumers, we assign price-anticipating consumers in a random uniform way to a period's first and second half and consider a correspondingly updated price history on a half period level.

Recall, the decision rules of all types of customers are not known to the agent.

4.2. Simulation, experimental setup and performance evaluation

In this section, we describe the simulation of our market framework and specify the environment, *i.e.* states, actions, rewards, and state transitions, to be used by RL algorithms (Section 4.2.1). Further, we provide all parameters to reproduce our experimental setup (Section 4.2.2) and explain the training procedure and performance evaluations (Section 4.2.3).

4.2.1. Simulation step algorithm and used observable state

The simulation of the sales events $i_t^{(k)}$ happens in an algorithm that is executed every timestep t, inhibiting all aspects of the vendor actions and the customer's buying behavior. Further, the state spaces are updated, and rewards are calculated each timestep. Fig. 1 presents an overview of the algorithm of one simulation step.

In our evaluation, the observable state of the agent is specified as follows. In the monopoly setup, the state s_t is given by the own last n_{last} period's prices, *i.e.*,

$$s_t := \left(H_t^{(n_{last})}\right) = \left(p_{t-1}^{(1)}, p_{t-2}^{(1)}, \dots, p_{t-n_{last}}^{(1)}\right). \tag{13}$$

In the duopoly setup, the agent's state s_t is given by the current price of the competitor ($p_t^{(2)}$) and the minimum of both competitors' prices for all half-periods within the last n_{last} periods, *i.e.*,

$$\begin{aligned} s_t &:= \left(p_t^{(2)}, H_t^{(n_{last})}\right) \\ &= \left(p_t^{(2)}, \min_{k=1,2}\left(p_{t-0.5}^{(k)}\right), \min_{k=1,2}\left(p_{t-1}^{(k)}\right), \dots, \min_{k=1,2}\left(p_{t-n_{last}}^{(k)}\right)\right). \end{aligned} \tag{14}$$

Hence, in the duopoly case the state s_t has $2 \cdot n_{last} + 1$ entries. The state transition in each step to s_{t+1} is straightforward, *i.e.*, the most recent minimum prices of the last two half periods are updated while the minimum prices of the two oldest half periods leave the state.

After the price offers $\vec{p}_t$ from the agent vendor and optional competitors are collected, we let constant amount of new customers i_{max} arrive. To determine how many customers arrive for each type, we randomly draw the absolute number of arriving customers per type, *i.e.*, $X_t^{(0)}, \dots, X_t^{(n_c)}$ out of i_{max}, using a multinomial distribution based on the probabilities $C(c)$, see (8). Hence, we have $E(X_t^{(c)}) = i_{max} \cdot C(c)$.

In the next step, we add the drawn number of waiting customers $\vec{w}_t$ for every customer type that is able to postpone purchases, cf. $c = 1, 2, 3$. Now, with $N_t^{(c)} = X_t^{(c)} + w_t^{(c)}$ we get the total number of customers of each type c eligible to buy an item at this timestep per type.

To update the competitor offers fairly, we split the simulation step into $n_{vendors}$ partial steps, where each vendor iteratively updates its offer, starting with the agent vendor. Therefore, we draw the belonging to partial steps for each customer and aggregate, yielding $N_t^{(c')}$. Starting with the iteration of the agent vendor, no competitor offers are updated. The next step is executed separately for every customer type. Using a defined customer behavior, we calculate the realized demand (in each partial step) based on the current price vector $\vec{p}$.

We use the demand probabilities $P_t^{(c,k)}(\vec{p}_t; H_t^{(h)})$ of each consumer type to multinomially draw the number of customers buying in this partial step. First, we do this for every customer; later, we accumulate the sales from all consumer types to calculate the number of sales of firm k for this partial step $i_t^{(k')}$. Here, we denote $Y_0^{(c')}, \dots, Y_{n_{vendors}}^{(c')}$ as the number of buying customers (for a partial step) per type c to finally determine $i_t^{(k')}$ for all firms k, *i.e.*,

$$i_t^{(k')} = \sum_{c'=0}^{n_c} Y_{k'}^{(c')}. \tag{15}$$

After that, we update the offer price of the first competitor, draw again, and repeat until every competitor was updated. In the end, we accumulate all $i_t^{(k')}$ to get $i_t^{(k)}$.

At the end of one simulation step, we calculate the reward (r_t) and update the state information ($\vec{s}_t$ including s_t). The seasonal component is updated via (1), while all not buying customers, *i.e.*, $Y_0^{(c)}$, enter their corresponding waiting pools, cf. $\vec{w}_t$, if they postpone their buy. Furthermore, the storage of last prices, cf. $H_t^{(h)}$, is updated to include the most recent offer a_t and drop the most previous out of the storage.

4.2.2. Reproducible experimental setup

For all our simulation runs, we use the parameters given in Table 1 if not stated otherwise.

4.2.3. Training and performance evaluation

Training: We define one training episode to last $T = 70$ timesteps (cf. 10 weeks) and set a limit of training episodes to 100 000. The number of training steps until convergence ranges from 10 000 training episodes in simple experiments to 75 000 in our most complex market environments, with markets of higher relative amounts of strategic customer behavior requiring longer training. After each training episode, the environment is set back to an initial state containing zero values for current prices, price histories, and waiting pools, cf. $\vec{s}_t := (0, \dots, 0)$.

Further, for all experiments, we use five independent training runs with different random seed values. All experiments were run on a MacBook Air with an Apple M1 processor and 8 GB RAM (runtime about 1 h per 20 000 episodes).

Performance Evaluation: By definition trained policies do only depend on the state considered in the infinite horizon MDP, see Section 3.1.4. Further, we use the learned policy without random effects (cf. exploration rate during training).

Since the vendor agent model bases its prediction on recent prices, in the beginning of an episode, offer prices do either not yet exist or

Table 1
Overview of the chosen model parameters for our experiments.

Variable	Explanation	Default
γ	Discount factor for one period	0.9999
$n_{seasons}$	Number of seasons (cycle length, types of periods, weekdays)	7
T	Number of timesteps per training episode (cf., e.g., 10 weeks)	70
a_{max}	Maximum offer price	10
A	Admissible prices (continuous)	$[0, a_{max}]$
$n_{vendors}$	Number of vendors (cf. monopoly (1) vs. duopoly case (2))	1 or 2
n_{last}	Number of last periods with $n_{vendors}$ prices each (cf. agent's state)	$n_{seasons}$
a_{floor}	Undercutting strategy price floor (competitor)	1
Δ	Undercutting difference (competitor)	1
n_c	Number of non-myopic customer types, $c = 0, 1, \dots, n_c$ (0 myopic)	3
i_{max}	Number of new arriving customers (aggregated over all types)	50
w_{max}	Maximum amount of waiting customers per type	1000
$u^{(0)}$	Utility of not buying (myopic customer)	1
α	Adjustment parameter for utility function u_t (myopic customer)	4
β_t	Level for utility u_t (seasonal demand case, $t = 0, \dots, n_{seasons} - 1$)	(4,6,7,3,6,5,7)
π_{remain}	Probability (per period) to stay in waiting pool (type $c = 1, 2, 3$)	0.95
π_{return}	Probability (per period) to visit the market (type $c = 1, 2, 3$)	0.95
δ	Number of periods for past reference (price-aware customer)	$n_{seasons} - 1$
h_{rel_ref}	Relative reference price threshold (price-aware customer)	0.9
WTP_{max}	Maximum willingness to pay (type $c = 2, 3$)	7
ϕ	Number of periods for forecasting input (price-anticipating cust.)	$2 \cdot n_{seasons}$
θ	Forecasting horizon in periods (price-anticipating customer)	$n_{seasons} - 1$
h_{rel_gap}	Relative price threshold (price-anticipating customer)	0.9

are distorted. Hence, evaluation runs include an initial phase of attunement. To obtain representative long-term performance results, we exclude this phase of attunement from the evaluation of a pricing policy (to lose the influence of the initial state). Further, due to the repeating dynamics of the $n_{seasons} = 7$ base periods in our infinite horizon model, cf. (1), the visited states and the associated realized price trajectories have a repeating cyclical structure. As this steady-state pattern goes on forever, it is sufficient to measure the performance for a finite selection of some of these steady-state cycles. To count only such steady-state cycles and to not include biases resulting from a (randomly) chosen starting state (phase of attunement), in our performance evaluation, we discard the first half of an episode (*e.g.* $T/2 = 70/2$ time steps, 5 cycles) – the time that is typically needed to reach the steady state. Then, for the remaining half of the episode, we measure the steady-state performance of the policy over $T/2$ time steps (5 cycles). Such evaluation runs are repeated.

For each experiment, we use 1000 evaluation runs (for all single training runs each) to analyze the trajectories of reward, offer price, and number of customer buys per timestep by visual inspection. Further, if the investigated customer behavior is able to postpone a purchase, we display the number of waiting customers. We average the results for each type of customer behavior, also considering the average sales price. In duopoly markets, we do this for both vendors and the total market.

Finally, in sensitivity analyses, we illustrate and discuss the impact of the share of a certain customer type by showing the average sales price and the number of customers buying with an increasing share of the non-myopic type, given myopic customers as the remaining base customer group. We again average results over 1000 evaluation runs (for 3 training runs) evaluating the agent's sales and, in case of a duopoly, include the competitor's sales.

4.3. Experiment 1: Myopic customers & seasonal demand (baseline model)

The first experiment serves as a baseline without strategic customer behavior to present our market simulation framework. We show the results for a monopoly (Section 4.3.1) and duopoly setup (Section 4.3.2) with customers having a season-dependent demand and compare the results to the optimal values in the monopoly case.

4.3.1. Monopoly setup

Fig. 2 presents the trajectories associated to the application of the optimized policy obtained after the training phase in a monopoly setup. To visualize the steady state, we refrain from the influence of initial states and show the second half of a simulation episode (*i.e.* 70/2 periods = 5 cycles). The agent sets the same prices every seven timesteps, *i.e.*, every cycle of seasons. Most customers buy for these prices, and low variations can be deducted from the stochastic demand behavior. Agent rewards follow the cyclical pattern. Fig. 3 further illustrates that, during training in a monopoly setup, the agent sets different offer prices for different seasons, converging after about 15 000 training episodes.

Comparison to optimal prices. For the setup with 100% myopic customers and seasonal demand Table 2 shows that the prices offered by the agent vendor are close to the optimal prices for each season, which in this basic setup are tractable under full information. We calculate the expected reward given a certain price and conclude that smaller differences in the offer price must not change the expected reward. We use the mean of all differences between actual and optimal prices and profits as performance metric, indicating how well the agent sets prices to maximize profits. In this experiment with a monopoly setup, the agent reaches 99.37% of the optimal prices and 99.96% of the optimal profits.

4.3.2. Duopoly setup

Fig. 4 shows trajectories of a duopoly setup against the undercutting competitor (7). We conclude that the agent finds cyclical offer prices for each season again. It is constantly undercut by Δ by its competitor. Agent and competitor mostly alternate in making many customers accept their offers, and variations are caused by stochastic customer behavior.

Nevertheless, the agent can gain a slight advantage when lowering its price offer as a response to a decreased demand, because of the higher competitor price offer in the following half timestep. Furthermore, Table 3 shows that the agent can achieve a higher total reward than the competitor, caused by higher offer prices on average. On the other hand, the competitor benefits each second half period because of its perfectly fitting undercut offer, adjusted to the agent's offer. Higher deviations from average sales and average offer prices of the competitor are caused by keeping the undercut price from the last timestep as well.

More generally, a duopoly setup can be advantageous for the customer base since the competitive offer might cause prices to decrease, causing a higher number of sales. Each vendor loses a significant amount of revenue and is forced to decrease price offers in comparison to a monopoly market.

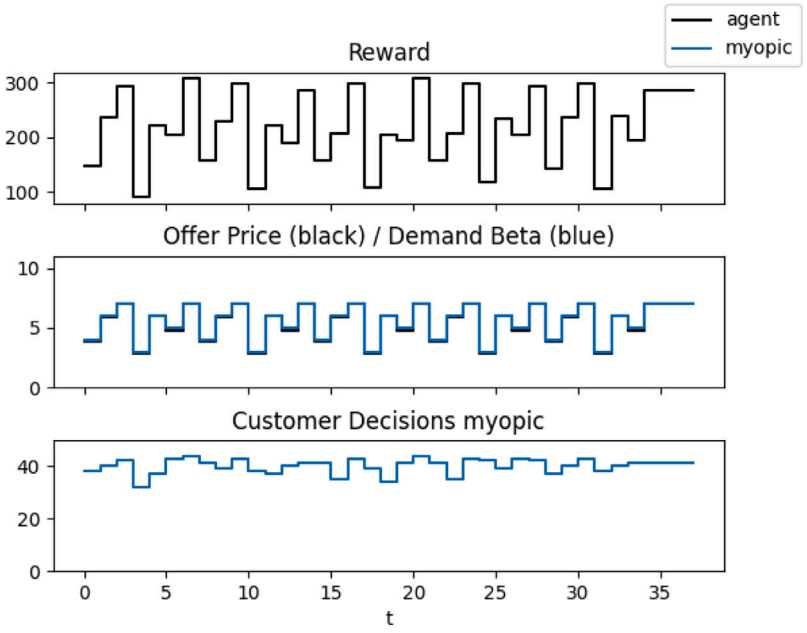

Fig. 2. 100% Myopic Customers in a Monopoly: Second half of a simulated episode with seasonal demand, *i.e.* rewards, offer prices, and realized sales over time. The policy was trained for 15 000 episodes and depends on the season, cf. $t \bmod n_{seasons} \in \{0, 1, \dots, 6\}$ (Section 4.3.1).

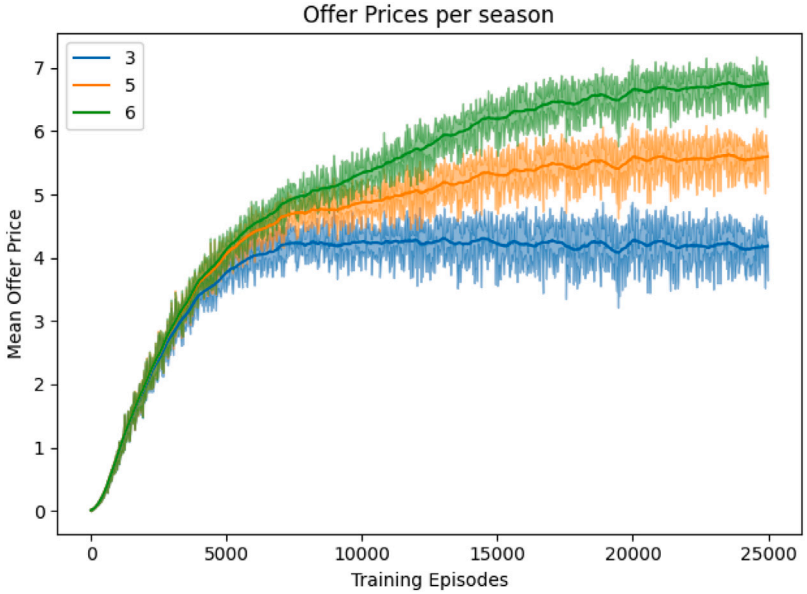

Fig. 3. Learned offer prices (and standard deviations) over 25 000 training episodes for three exemplary seasons with low, medium, and high demand (cf. $\beta_3 = 3$, $\beta_5 = 5$, $\beta_6 = 7$) in a monopoly setup (Section 4.3.1), averaged over 5 independent training runs.

Main findings 1. This experiment demonstrates the correct modeling and implementation of a seasonally dependent demand causing cyclic price patterns in a monopoly and a duopoly setup. We compare the results of the learned policy of our vendor agent to a tractable optimal solution in the monopoly case and conclude that when applying PPO we are able to reach this optimal solution after less than 15 000 training episodes.

4.4. Experiment 2: Recurring customers

Experiment 2 studies the customer behavior of type $c = 1$ that is strategic to a small degree and follows simple rule-based mechanisms. We aim to observe whether minor strategic considerations, even marginally as postponing a purchase, influence the learned pricing policy.

4.4.1. Monopoly case

Fig. 5 shows simulated trajectories of a market with only recurring customers in a monopoly setup. It is visible that the agent vendor can benefit from the customers' ability to wait, caused by the seasonally dependent demand. Contrary to the customers' intended behavior of waiting for the lowest offer price, the agent finds the seasons with highest demand and sets its offer price of all seasons to the optimal price for those specific seasons. This causes a behavior where customers do not buy during the seasons of lower demand. They wait until the seasons with the highest demand, find an offer price that fits, and accept the high offer price.

Table 2
Comparison of learned vs. optimal prices and associated rewards for all 7 seasons (cf. 0-6) and in total (∑) for 100% myopic customers with seasonal demand in a monopoly setup. The agent is able to learn a near-optimal policy after 15 000 training episodes (Section 4.3.1). Results are averaged over 5 independent training runs and 1000 evaluation runs each.

	Metric	Demand Beta (β_t)					∑
		3	4	5	6	7	
	Season	3	0	5	1 & 4[a]	2 & 6[a]	0–6
Actual	Offer Price	2.77	3.85	4.77	5.98	6.97	
	Reward per Customer	2.03	2.94	3.83	4.75	5.65	29.60
	Reward per Timestep	101.70	147.14	191.29	237.29	282.35	1 480.21
	Reward per half Episode	508.48	735.70	956.46	1 186.45	1 411.74	7 401.06
Optimal	Offer Price	2.76	3.85	4.92	5.97	7.02	
	Reward per Customer	2.03	2.94	3.85	4.75	5.65	29.62
	Reward per Timestep	101.73	147.14	192.38	237.44	282.35	1 480.83
	Reward per half Episode	508.65	735.70	961.90	1 187.20	1 411.75	7 404.15

[a] The values for seasons 1 & 4 and seasons 2 & 6 (same β_t) are similar and have been averaged.

Fig. 4. 100% Myopic Customers in a Duopoly against the undercutting competitor: Second half of a simulated episode with seasonal demand, *i.e.*, both firms' rewards, offer prices, and sales over time. The agent's policy was trained for 15 000 training episodes (Section 4.3.2).

Table 3
100% Myopic Customers in Monopoly vs. Duopoly Setups: Performance metrics of Experiment 1 with seasonal demand, i.e., offer prices, sales prices, number of sales, and rewards. Results are averaged over 5 independent training runs and 1000 evaluation runs each. ∅ denotes the mean average and ∑ denotes the total sum per episode (Section 4.3.2).

	Metric	Monopoly	Duopoly
Agent	∅ Offer Price	5.25	4.92
	∅ Sales Price	5.27	4.71
	∑ Customers Buying	1 401.88	705.44
	∑ Revenue	7 382.09	3 319.40
Comp.	∅ Offer Price	–	3.92
	∅ Sales Price	–	3.57
	∑ Customers Buying	–	883.28
	∑ Revenue	–	3 151.55
Total	∅ Sales Price	5.27	4.07
	∑ Customers Buying	1 401.88	1 588.71
	∑ Revenue	7 382.09	6 470.95

4.4.2. Duopoly case

In a duopoly market, see Fig. 6, prices follow a cyclical pattern of the defined week length. Prices are set to a high price near $a_{max} = 10$, avoiding that recurring customers accept the offer of the competitor, cf. causing them to postpone their buy. In the first half of a timestep following the low demand, the agent achieves high reward caused by the not updated competitor offer. After undercutting this offer, the competitor profits in the second half. Fewer customers wait in the duopoly market.

Table 4 presents a numerical evaluation of the recurring customer. The agent exploiting the recurring behavior can achieve a sales price of 7.00 in the monopoly setup and has many customers accepting this offer. Rewards are higher than in a market with myopic customers having the same seasonal demand but not being able to wait. In the duopoly market, this effect fades, by the competitor causing a cyclical price pattern with decreasing offer prices.

Counteracting this effect, we also evaluated the case with 90% myopic customers. The resulting trajectories are similar to those of Experiment 1, cf. Figs. 2 and 4, retaining the cyclical price pattern caused by different demands per season. The numerical evaluation in Table 4 shows that the remaining 10% recurring customers are still performing slightly worse than their myopic counterparts, due to their identical demand model. Overall, the market simulation with a small share of recurring customers mostly equals one with only myopic customers.

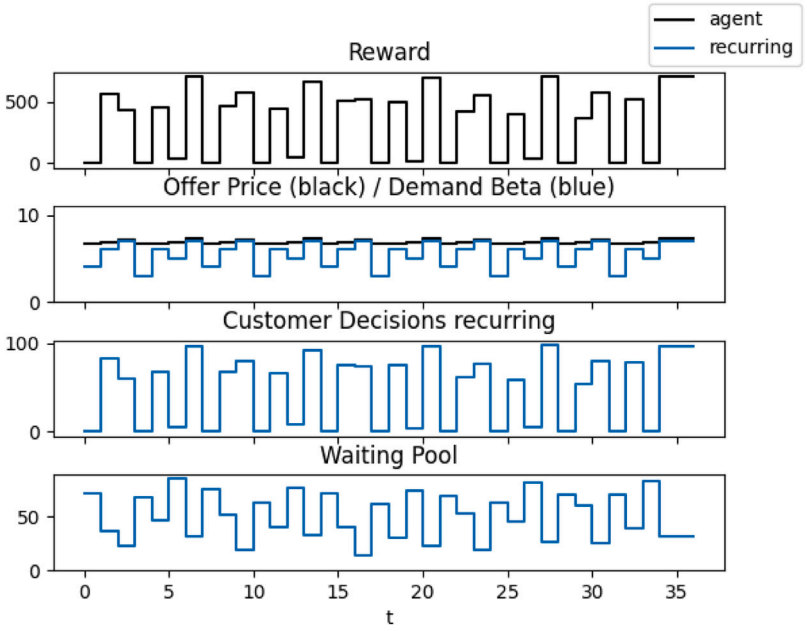

Fig. 5. 100% Recurring Customers in a Monopoly: Second half of an episode with seasonal demand, *i.e.*, rewards, offer prices, sales, and waiting pool (after 10 000 training episodes). The agent reacts to the ability to wait by not lowering prices for low demand seasons (Section 4.4.1).

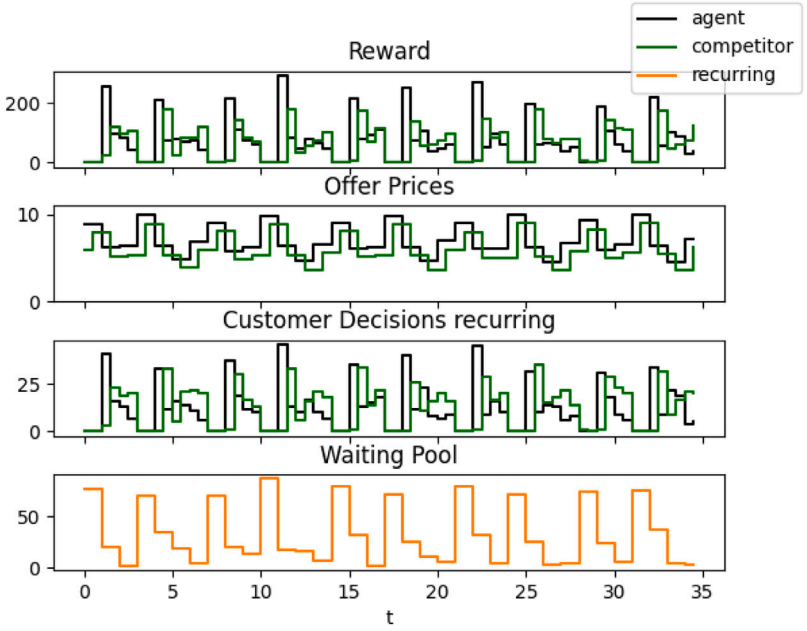

Fig. 6. 100% Recurring Customers in a Duopoly against the undercutting competitor: Second half of a simulated episode with seasonal demand, *i.e.*, both firms' rewards, offer prices, sales, and the waiting pool size over time (results after 20 000 training episodes) (Section 4.4.2).

4.4.3. Sensitivity analysis: Impact of the share of recurring customers

In order to analyze the impact of the share of recurring customers, we run simulations for different compilations of myopic and recurring customers. We increase the relative number of recurring customers by ten percent until no myopic customers are left. Fig. 7 shows the average sales price per customer and the normalized number of customers buying in both market setups over 3 independent training runs with each 1000 simulation runs.

The evaluation of the average sales prices supports our conclusion above. In a monopoly market, the agent increases prices, contrarily letting customers wait for higher prices as more recurring customers enter the monopoly market. With increasing prices, the number of sales decreases more for myopic customers. This holds true only to some extent in the duopoly market. Sales prices increase from a lower averaged price with only myopic customers, causing a decrease in myopic sales after a share of 60% recurring customers. In both markets, more recurring customers buy caused by the chance to return to the market after not accepting the first offer.

Main findings 2. Experiment 2 presents a market simulation with a customer behavior being able to postpone purchases. We identify that this ability does not guarantee a better performance, bringing along the ability to be exploited by a vendor agent. Such an exploitation can be circumvented to a small degree by consulting a competitive vendor, or

Table 4
10% & 100% Recurring Customers in Monopoly vs. Duopoly Setups: Performance metrics of Experiment 2 with seasonal demand, i.e., offer and sales prices, realized sales, and rewards. Results are averaged over 5 independent training runs and 1000 evaluation runs each. ∅ denotes the mean average, ∑ denotes the total sum per episode (Sections 4.4.1–4.4.2).

	Metric	Monopoly			Duopoly		
			Combination of			Combination of	
		1.0 rec.	0.9 myo.	0.1 rec.	1.0 rec.	0.9 myo.	0.1 rec.
Agent	∅ Offer Price	6.84	5.19		6.46	4.86	
	∅ Sales Price	7.00	5.24	5.25	5.67	4.67	4.68
	∑ Cust. Buying	1 567.33	1 261.38	170.60	856.77	636.38	77.41
	∑ Revenue	10 975.14	6 612.07	896.18	4 860.02	2 971.39	362.58
Comp.	∅ Offer Price	–	–	–	5.45	3.86	
	∅ Sales Price	–	–	–	4.49	3.53	3.53
	∑ Cust. Buying	–	–	–	835.03	793.50	96.73
	∑ Revenue	–	–	–	3 742.67	2 800.52	341.22
Total	∅ Sales Price	7.00	5.24	5.25	5.09	4.04	4.04
	∑ Cust. Buying	1 567.33	1 261.38	170.60	1 691.80	1 429.88	174.14
	∑ Revenue	10 975.14	6 612.07	896.18	8 602.69	5 771.91	703.80

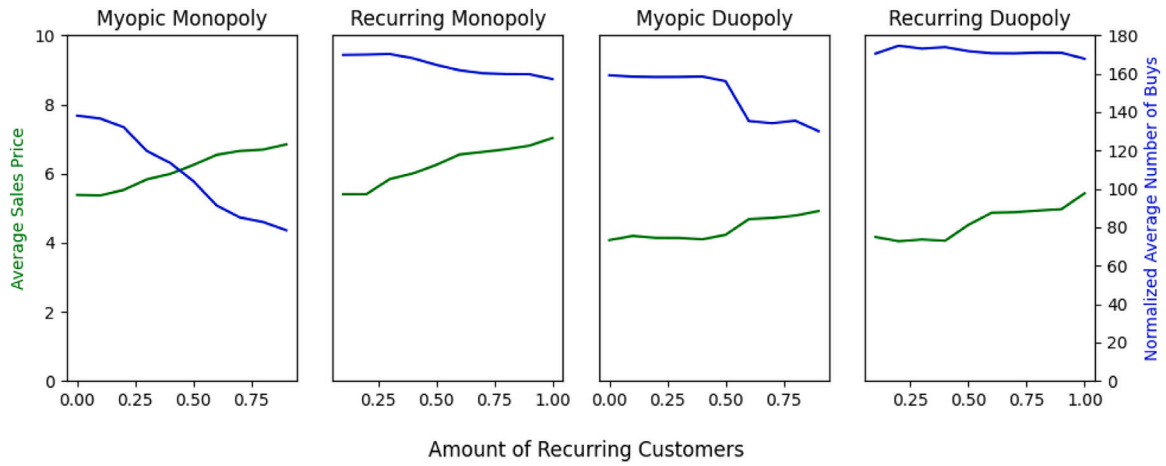

Fig. 7. Impact of the Share of Recurring Customers (with seasonal demand) in % (x-axis) if the remaining share of customers is myopic: Average sales prices and number of sales (normalized) for both types, *i.e.*, the recurring customers and the myopic customers, in the monopoly and the duopoly setup (each after 15 000–20 000 training episodes). Results are averaged over 3 independent training runs and 1000 evaluation runs each (Section 4.4.3).

avoided by using a sufficiently large group of myopic customers causing the optimal pricing policy to restore the seasonal price pattern.

Independent of the ability to be exploited, our modeling of the recurring customer cannot achieve significantly higher consumer rent than their myopic counterparts. This is caused by having the same demand function and no rule explicitly aiming at a lower price but instead just not accepting too high prices.

4.5. Experiment 3: Price-aware customers

In contrast to the recurring customer, our modeling of a price-aware customer is able to store recent prices and thereby to decide based on historical reference data. This modeling does not predict future prices but waits until the current offer price is lower than past reference prices.

4.5.1. Monopoly case

Fig. 8 presents trajectories of a simulation episode of such a market with only price-aware customers in a monopoly market. A pattern of prices can be observed where the agent sets the highest possible price for a series of timesteps as a preparation to decrease the offer afterwards. In the timesteps following this series of high price offers, the agent constantly decreases prices such that customers accept the offers, until reaches a very low price, and, finally, increases prices again.

The chance of customers being lost during a series of not purchasing, given by π_{remain} and π_{return} causes the notable surpassing of WTP_{max} of the offer price to be beneficial. Instead of four times setting a price of ten followed by one offer for 6.99, the agent does not lose waiting customers by further decreasing prices. That is why the length of one price cycle is just dependent on the possible price range in this scenario.

The numerical evaluation in Table 5 shows a large difference between average offer prices and average sales prices in the market with 100% price-aware customers. When evaluating a market with a small share of price-aware customers, they are able to outperform their myopic counterparts, achieving an average sales price of 51.39% of the myopic with an ordinary number of purchases.

4.5.2. Duopoly case

In a duopoly market with 100% price-aware customers, the agent and the competitor constantly undercut each other until the competitor offers its floor price $\alpha_{floor} = 1$. Because of the agent's ability to further surpass the competitors' price, all customers accept the very low offer of the agent, cf. Table 5.

We show a market with 10% price-aware customers and 90% myopic customers in Fig. 9. The cyclic price pattern for the myopic customers is restored. The price-aware customer always accepts the offer of the competitor in the second half of the season with minimum demand. Since the agent does not make profit with price-aware customers in a duopoly setup at all, the pricing pattern is not optimized

Fig. 8. 100% Price-Aware Customers (with $WTP_{max} = 7$) in a Monopoly: Second half of an episode, *i.e.*, rewards, offer prices, sales, and waiting pool (after 60 000 training episodes). The "Price-aware Demand" indicates the highest price the customer would accept in t (Section 4.5.1).

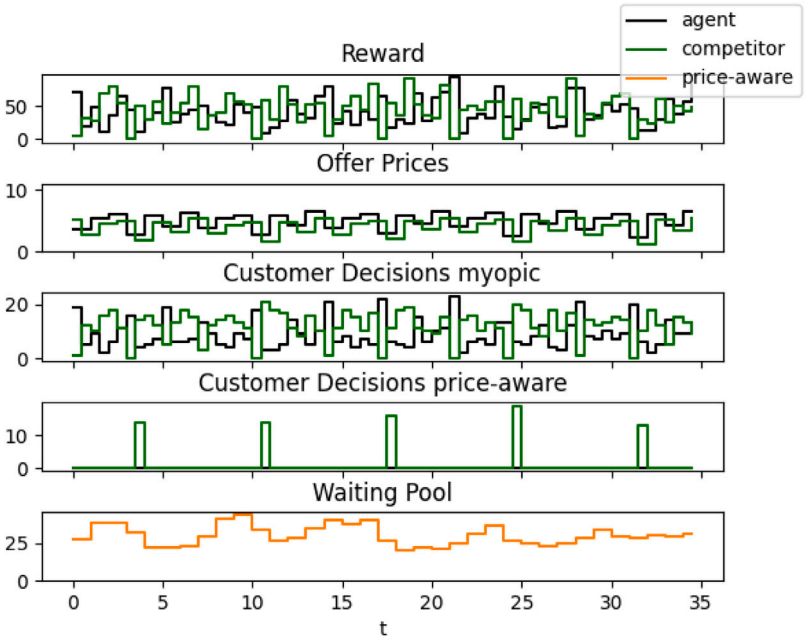

Fig. 9. 10% Price-Aware Customers (with $WTP_{max} = 7$) in a Duopoly against the undercutting competitor (7): Second half of a simulated episode, *i.e.*, both firms' rewards, offer prices, sales, and the waiting pool size over time (after 17 500 training episodes). The 90% myopic customers lead to a pricing pattern than can be exploited by price-aware customers (Section 4.5.2).

towards exploiting them. That is why the waiting pool shows a less cyclical trajectory. Only a smaller number of price-aware customers buys the product for 42.99% of the average price of myopic customers, cf. Table 5.

4.5.3. *Sensitivity analysis: Impact of the share of price-aware customers*

Studying different relative amounts of customers, Fig. 10 presents the average sales prices and normalized average number of sales in both market settings. In comparison to the recurring customer, it can be observed that the price-aware is achieving lower sales prices than the myopic base customer group introducing the seasonal demand. The price-aware customer benefits highly from the agent vendor offering near-optimal prices for the myopic base group until there are more than 50% price-aware customers in a monopoly market. After this point, the agent adopts prices more towards maximizing profits from price-aware customers, resulting in lower average sales prices of the myopic and higher prices for the price-aware behavior. Despite an increase in sales prices, there are more price-aware customers buying.

In a duopoly setup, the realized sales price as a combination of agent and competitor sales is increased more steep after a higher share of price-aware customers participates in the market. This yields an effect of decreasing number of sales of the myopic customers with increasing

Table 5
10% & 100% Price-Aware Customers (with $WTP_{max} = 7$) in Monopoly vs. Duopoly Setups: Performance metrics of Experiment 3, i.e., offer prices, sales prices, realized sales, and rewards. Results are averaged over 5 independent training runs and 1000 evaluation runs each. ∅ denotes the mean average, ∑ is the total sum per episode. The 90% myopic customers lead to a pricing pattern that is exploited by price-aware customers (Sections 4.5.1–4.5.2).

	Metric	Monopoly			Duopoly		
			Combination of			Combination of	
		1.0 p.-aw.	0.9 myo.	0.1 p.-aw.	1.0 p.-aw.	0.9 myo.	0.1 p.-aw.
Agent	∅ Offer Price	6.75	5.32		1.22	4.91	
	∅ Sales Price	5.43	5.37	2.76	0.69	4.70	4.00
	∑ Cust. Buying	1 470.46	1 239.77	131.60	1 179.54	635.18	0.01
	∑ Revenue	8 005.33	6 655.29	363.15	808.65	2 983.33	0.04
Comp.	∅ Offer Price	–	–	–	1.01	3.91	
	∅ Sales Price	–	–	–	1.00	3.56	1.75
	∑ Cust. Buying	–	–	–	0.10	795.52	86.62
	∑ Revenue	–	–	–	0.10	2 834.01	151.99
Total	∅ Sales Price	5.43	5.37	2.76	0.69	4.07	1.75
	∑ Cust. Buying	1 470.46	1 239.77	131.60	1 179.64	1 430.70	86.83
	∑ Revenue	8 005.33	6 655.29	363.15	808.75	5 817.34	152.05

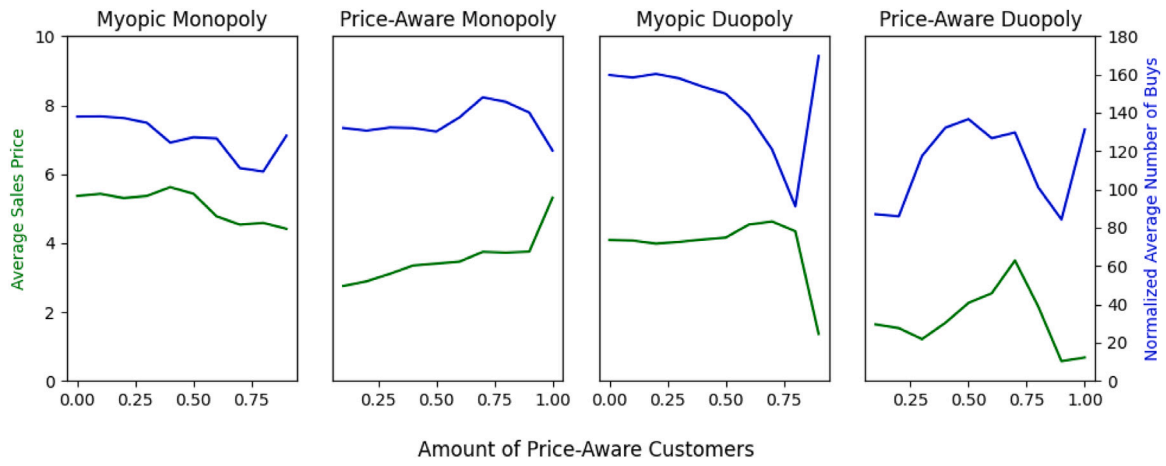

Fig. 10. Impact of the Share of Price-Aware Customers ($WTP_{max} = 7$) in % (x-axis) if the remaining share of customers is myopic: Average sales prices and number of sales (normalized) for both types, *i.e.*, the price-aware customers and the myopic customers, in the monopoly and the duopoly setup (each after 15 000–60 000 training episodes). Results are averaged over 3 independent training runs and 1000 evaluation runs each (Section 4.5.3).

share of the price-aware type, until a price-race-to-the-bottom results from over 80% price-aware customers. This causes high numbers of sales with low average sales prices for both types.

Main findings 3. This experiment presents market dynamics when including strategic customer behavior, being able to observe past prices. Despite its inherent ability to be exploited in a monopoly market without a base customer group having seasonal demand, the price-aware buying behavior benefits highly in a market having cyclical offer price patterns. In monopoly and duopoly markets with 90% myopic customers, they achieve discounts of 50%–60% in comparison to non-strategic behavior.

4.6. Experiment 4: Price-anticipating customers

In the last experiment, we study counteracting the price-anticipating customer behavior, using an auto-regression to predict future prices, thereby being forward-oriented instead of solely observing past prices.

4.6.1. Monopoly case

Fig. 11 presents simulated trajectories for a monopoly market with only price-anticipating customers. In contrast to Experiment 3, the agent is not able to exploit the anticipating behavior just by setting a series of prices to a_{max}, but rather offering a sinusoidal price trajectory. The predictions of the anticipating customer are not exactly matching the offer prices, nevertheless, the anticipating customers achieve a better average sales price, *i.e.* 5.83, than the average offer price of 6.63.

In a market with 90% myopic customers with seasonal demand and 10% price-anticipating customers, we find that the price-anticipating customers can gain a competitive advantage over the myopic customers, paying 49.04% of the average sales price of the myopic customers. A sufficiently high amount of price-anticipating customers are purchasing and achieving a better average sales price than the price-aware behavior, cf. Table 6.

4.6.2. Duopoly case

In a duopoly market, the price-anticipating customer predicts the minimum of two price trajectories by always taking the minimum offer price as input to the auto-regression as described in Section 3.2.4. This again causes customer behavior where anticipating customers do not accept the higher price offer, similar to the price-aware behavior in Section 4.5.2. Further, the undercutting competitor (7) always offers a lower price, except when the price floor is reached.

We studied a duopoly market with price-anticipating customers only and a duopoly market with 90% myopic customers as a base group. While the composition of a market with two vendors and only price-anticipating customers results in a pricing policy of offer prices near the price floor of the undercutting competitor in our modeling, Fig. 12 shows the simulation trajectories of a market with 10% price-anticipating and 90% myopic customers. It can be observed that the anticipating customers accept minimal price offers from the competitor, while remaining market dynamics follow the outcome of a duopoly market with only myopic customers.

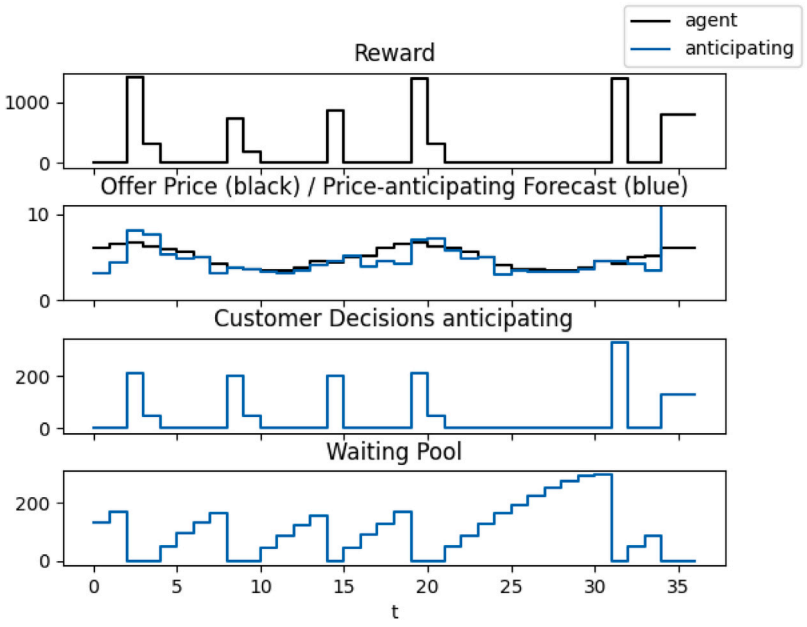

Fig. 11. 100% Price-Anticipating customers (with $WTP_{max} = 7$) in a Monopoly: Second half of an episode, *i.e.*, rewards, offer prices, sales, and waiting pool (results after 75 000 training episodes). The "Price-anticipating Forecast" indicates the forecast in $t + 1$ for t (Section 4.6.1).

Fig. 12. 10% Price-Anticipating Customers (with $WTP_{max} = 7$) in a Duopoly against the undercutting competitor (7): Second half of a simulated episode, *i.e.*, both firms' rewards, offer prices, sales, and the waiting pool size (after 25 000 training episodes). The 90% myopic customers lead to a pricing pattern that can be exploited by anticipating customers (Section 4.6.2).

Analyzing the metrics in Table 6, it can be concluded that the competitor causes offer prices to decrease, resulting in another 42.20% discount of the already low average sales price in a monopoly market with 10% price-anticipating customers. A sufficient number of purchases is realized, all accepting the competitor offer.

4.6.3. Sensitivity analysis: Impact of the share of price-anticipating customers

Investigating the impact of the share of price-anticipating customers in a monopoly market with myopic customers having seasonally dependent demand as a base group, we can observe that price-anticipating customers are able to benefit only if they are a minority in the simulated market, see Fig. 13. If there are more than 30% price-anticipating customers in our simulation market, this type of customer cannot outperform the myopic customers regarding consumer rent. In some configurations, they even pay higher average sales prices.

In a duopoly market, sales prices are increasing with rising share of anticipating customers until 70% while normalized myopic sales decrease. After a certain point the average sales price for both customer types drop, resetting the number of myopic sales to the prior level that was realized with a large share of myopic customers.

Main findings 4. We studied the price-anticipating customer behavior using an auto-regressive forecast in a monopoly and duopoly setup.

Table 6
10% & 100% Price-Anticipating Customers ($WTP_{max} = 7$) in Monopoly vs. Duopoly Setups: Performance metrics of Experiment 4, i.e., offer/sales prices, realized sales, and rewards. Results are averaged over 5 independent training runs and 1000 evaluation runs each. ∅ denotes the mean average, ∑ is the total sum per episode. The 90% myopic customers lead to a cyclic pricing pattern that is exploited by anticipating customers (Sections 4.6.1–4.6.2).

	Metric	Monopoly			Duopoly		
			Combination of			Combination of	
		1.0 antic.	0.9 myo.	0.1 antic.	1.0 antic.	0.9 myo.	0.1 antic.
Agent	∅ Offer Price	6.63	5.17		1.32	4.88	
	∅ Sales Price	5.83	5.20	2.55	0.79	4.65	4.60
	∑ Cust. Buying	1 663.41	1 266.69	79.69	1 175.89	641.79	0.02
	∑ Revenue	9 686.40	6 589.64	203.28	936.24	2 981.20	0.092
Comp.	∅ Offer Price	–	–	–	1.00	3.87	
	∅ Sales Price	–	–	–	1.00	3.53	1.46
	∑ Cust. Buying	–	–	–	0.12	793.18	69.17
	∑ Revenue	–	–	–	0.12	2 797.85	102.71
Total	∅ Sales Price	5.83	5.20	2.55	0.89	4.03	1.49
	∑ Cust. Buying	1 663.41	1 266.69	79.69	1 176.02	1 434.97	69.19
	∑ Revenue	9 686.40	6 589.64	203.28	936.36	5 779.05	102.80

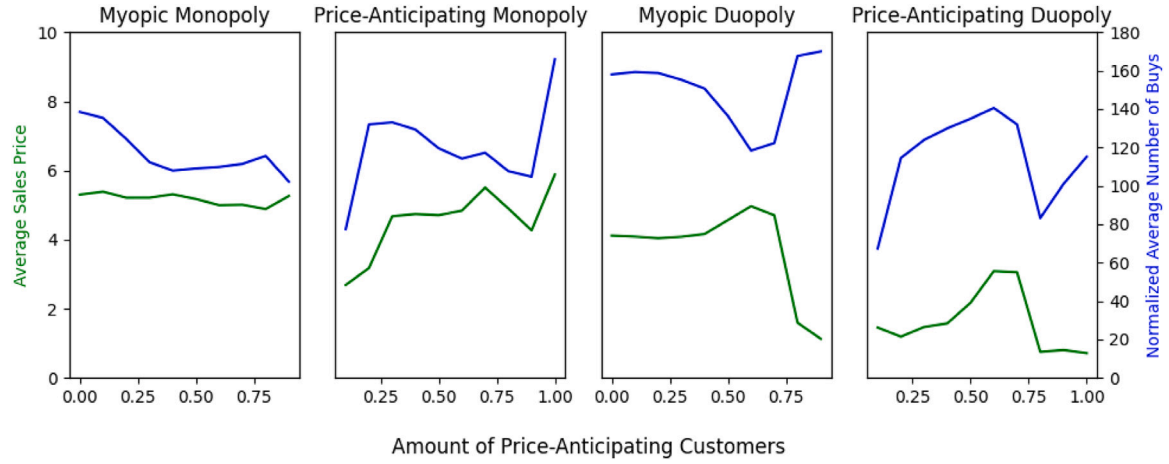

Fig. 13. Impact of the Share of Price-Anticipating Customers ($WTP_{max} = 7$) in % (x-axis) if the remaining share of customers is myopic: Average sales prices and number of sales (normalized) for both types, *i.e.*, the price-anticipating customers and the myopic customers, in the monopoly and the duopoly setup (each after 15 000–75 000 training episodes). Results are averaged over 3 independent training runs and 1000 evaluation runs each (Section 4.6.3).

With the ability to include the impact of single price decisions on consumers' price anticipations the agent accomplishes to deal with an unknown share of price-anticipating customers among myopic consumers. We find that anticipating prices and adopting the buying behavior accordingly causes an increase in consumer rent, which can be further heightened by a competitor offering lower prices. In total, this behavior outperforms the price-aware behavior in our modeling. In most of our studied market setups, the price-anticipating customers are able to identify the minimum offer and postpone their buy when given a higher option, leaving fewer opportunity to be exploited by a vendor adopting its pricing policy.

5. Discussion

In this section, we summarize our main results and infer managerial insights (Section 5.1). Furthermore, we present limitations of our study (Section 5.2) and examine future research opportunities (Section 5.3).

5.1. Summary of results and managerial insights

In the following, we summarize our insights into how a merchant can counteract different strategic consumer behaviors in different situations. Finally, we also discuss which consumer behavior performs best in different market scenarios.

5.1.1. Counteracting myopic consumer behavior

Experiment 1 showed a monopoly and duopoly market containing myopic customers with unknown stochastic demand behavior, depending on the current season. In this simple baseline setup, the MDP is tractable under full information, allowing the calculation of the optimal policy. This allowed to verify that the RL agent is able to learn near-optimal policies without prior knowledge. The optimal solution involves adjusting offer prices per type of season. Prices can differ a lot and following the seasonal demand intensity of the (myopic) consumers.

In case of competition (against an undercutting merchant) – as expected – we overall observe lower prices and reduced rewards. Now, the agent has to balance between undercutting the competitor to increase profits and raising prices again to avoid a price race to a less rewarding price level at the bottom. This is achieved by increasing prices in periods of low demand, such that the agent offers the better price in periods with high possible reward. We find that in both market scenarios the required training time is modest.

5.1.2. Counteracting patient consumer behavior

Next, in Experiment 2, we studied a recurring customer behavior, also completely unknown to our vendor agent. In case of a decision against purchasing for a given price offer, this type might return to the market at a later period. In line with earlier studies, see [32], the waiting ability serves vendors to counteract their customers by

introducing higher prices at times of low demand. This causes them to visit the market in future periods of high demand, reducing their consumer rent in case of a purchase. That is why in markets with high share of recurring customers, the vendor agent offers stable prices at the highest demand level, preventing customers to purchase smaller offers.

Myopic customers experience a slightly decreasing consumer rent with an increasing share of recurring customers, caused by increasing offer prices in periods of low demand. On the other hand, the vendor agent benefits from exploiting the recurring behavior. This effect is less pronounced in a competitive market as the agent additionally has to deal with the interfering effect of being undercut, similar to Experiment 1.

5.1.3. Counteracting reference price-based consumer behavior

In Experiment 3, we considered recurring customers with a price-aware buying behavior, suitable to recognize and wait for small prices. This behavior causes increased consumer rent in market setups with a large base group of myopic customers. However, in case the share of price-aware consumers surpasses a certain threshold, the agents focuses on maximizing revenue from this type. Price-aware consumers are counteracted by first offering constant prices at a very high level, which generates a high reference price and a swelling waiting pool. Second, the price is dropped to the maximum willingness to pay such that a large number of price-aware customers within the waiting pool purchase. In few subsequent periods prices are dropped again (cf. skimming policies), which can be explained as follows. Raising the price directly and starting another phase of restoring the reference price results in losing customers that are still in the waiting pool (they return to the market with a probability smaller than 1 and leave the pool with a positive probability). Below the threshold (in our setting) of about 60% price-aware customers, it is not worth investing in high reference prices as the agent loses sales from the myopic customers.

Price-aware customers perform noticeably better than myopic customers in monopoly and duopoly market setups as long as there is a sufficiently large myopic base group. However, in this case, myopic customers are not affected by the strategic customer group entering the market. The agent suffers from a decent share of price-aware customers while it benefits from a large share in monopoly setups. In duopoly setups, the agent suffers even more from price-aware customers as the agent cannot form the reference price alone. Under competition, the agent either sets really low offer prices to sell to a large share of price-aware customers or focuses on the more rewarding myopic customers, dependent on the relative combination, *i.e.*, the associated expected total reward.

5.1.4. Counteracting price-anticipating consumer behavior

Finally, in Experiment 4, we studied a multi-period price-anticipating behavior, yielding buying decisions at the expected minimum price offer of a price trajectory. We find that this behavior is effective to recognize and wait for small prices, given a number of last offer prices as input. As long as the share of myopic customers is large, this customer type is able to benefit even more than price-aware customers. In scenarios with a small share of strategic customers, the waiting pool accumulated until the time of the predicted price minimum, resulting in many positive customer decisions at that period. Again, with a large share of anticipating customers, the vendor agent counteracts this behavior by setting long price trends aiming to find conditions fulfilling the strategic customer decision rule at high prices close to their maximum WTP.

We find that price-anticipating customers outperform all other behaviors in monopoly and duopoly markets with a large myopic base group. The vendor agent suffers from a small share of price-anticipating customers and benefits from a large one in a monopoly setup. In a duopoly setup with many price-anticipating customers, prices oscillate near the price floor of the competitor strategy. Note, that the agent cannot form the price history – that is crucial for price forecasts and in turn, for future demand – alone. Further, the undercutting competitor always sets the lowest price within a cycle, which besides, is also anticipated by strategic consumers. In a duopoly with few price-anticipating customers, the agent focuses on maximizing reward from the myopic base group, having to consider the competitor behavior as well.

5.1.5. Comparison of the effectiveness of different strategic consumer behaviors

Our experiments also allow to compare the performance of different buying behaviors from a consumer perspective. Next, for small as well as large sizes of the myopic base group, we summarize which behavior performed best with regard to consumer rent. We distinguish monopoly and duopoly setups.

- In a monopoly with 90% of myopic consumers, for the remaining 10% the price-anticipating behavior performed best.
- In a monopoly with 10% of myopic consumers, for the remaining 90% the price-aware behavior performed best.
- In a duopoly with 90% of myopic consumers, for the remaining 10% the price-anticipating behavior performed best.
- In a duopoly with 10% of myopic consumers, for the remaining 90% the price-aware behavior performed best.

Overall, we find that strategic customer behavior exceeding the ability to postpone a purchase by a forward or backward-looking element is beneficial. In monopoly and duopoly markets with 10% price-aware or price-anticipating customers, they achieve discounts of 50%–60% compared to a myopic base group.

Naturally, our results obtained from the considered numerical examples may not hold in general. Nevertheless, the proposed framework remains a useful tool to study the effects of strategic interaction of competing vendors and strategic customers.

5.2. Limitations

The limitations of our model are the following:

The lack of testing different RL algorithms can be seen as a limitation. Also, results are to some degree stochastic and a larger number of runs would have to be used to quantify mean values and their standard deviations accurately. Further, also a thorough hyperparameter tuning for PPO could have been performed.

In the existing framework, alternative rule-based competitor strategies or oligopoly setups could have been evaluated. The equidistant timing of both competitors to adapt their prices could be generalized or randomized. Moreover, letting also the competitor apply self-learning strategies is worth investigating. However, in such cases the dynamics of the environment change if competitors change their behavior (cf. Markov property).

Exploring other demand setups or decision rules might bring about different results and could be worth exploring in future research:

First, consumer arrivals could also be set up differently. Our approach of drawing the customer types and their buying decisions multinomially (with a constant number of new arriving customers) out of a demand probability yields valid results but could also be generalized. Also, additional firm-specific (*e.g.*, sustainability) features could be included [49].

Second, adapting or randomizing hyperparameters such as the thresholds for the maximum willingness-to-pay or the relative comparisons, as well as probabilities to wait and to return would be interesting. Also, the demand parameter β_t could be chosen differently in further examples. Further, variations of different price history lengths might lead to other phenomena.

Third, also various mixtures of shares of different customers of certain type [50] could be investigated regarding the firm's pricing policy and rewards as well as the mutual impact on the different customer types' performances.

Fourth, the modeling of a price-anticipating customer could as well be done with other regressive forecasts. The strengths of the model to be able to endogenize the impact of such forecasts, however, requires the history lengths of the sellers and vendors to be of somewhat similar lengths (cf. size of the state space).

Another simplification of our model is the absence of inventory considerations. In this regard, considering problem versions with remaining inventory levels in finite horizon settings or joint ordering decisions in infinite horizon settings seem possible. In our model, the state space as well as the action space would have to be extended by single dimensions. While this seems tractable, the complexity of the model increases and in turn, the required amount of training steps might increase as well.

Finally, the direct application of the model in practice is not possible as thousands of periods of exploration cannot be done in an online manner. Instead, it is required to define a suitable environment that allows to pretrain agents effectively. Here, the data-driven calibration of auxiliary training environments or digital twins is a promising opportunity (but this is not focus of the paper).

5.3. Future research

Our work poses a step towards the evaluation of complex market simulations and the learning of near-optimal pricing policies counteracting strategic customer behavior. Despite the required effort to investigate different buying strategies, there is a research gap for more complex market situations combined with customer behavior that is strategic to a high degree. For instance, the influence of more than two vendors in an oligopoly setup might bring along different effects on optimized actions for all market participants. Furthermore, inventory constraints, particularly in finite horizon settings, could be researched regarding the strategy of anticipating customers. Note, in this context, the consideration of a potential run out (cf. product availability) or a depreciation of the product value could be taken into account by strategic consumers. Other market characteristics could be altered as well, such as using continuous time intervals, allowing only markdown- or markup pricing policies, or varying information settings.

In addition, the usage of RL to optimize the decision-making *on the customer side* might also be an interesting research field. Training multiple agents on both vendor and customer side can enable the identification of optimal policies in more complex market settings. However, it is challenging to fulfill the Markov property of events being independent of historical actions in the modeling of such a market.

Finally, the applicability of this framework in practice has to be researched. We demonstrate that non-myopic customer behaviors can be modeled differently and examine essential factors for predicting a buying policy based on past prices. A vendor was able to counteract some of these strategies. In practice, it has to be discovered whether there is sufficient data to create and use a digital twin environment like our market simulation framework. [51], for instance, show how to use estimations of demand probabilities and competitor reactions in recommerce markets by using linear regressions based on historical data to define a digital twin, which is used for pretraining before applying the learned policy in an unobservable market environment. Similar approaches could be used for our models as well.

6. Conclusion

In this paper, we have studied how to use RL techniques to approximate equilibria between price-optimizing vendors and different kinds of backward- and forward-looking customers. Besides monopoly markets we also studies duopoly markets against rule-based competitors. Note, such equilibria of vendor and consumer strategies – that mutually influence each other – can only be identified in dynamic games if the vendor's state accounts for all price information that is relevant to the consumers' decisions rules. Based on a concise modeling of consumer behaviors and a corresponding counterpart within a tractable state space on the vendor's side we are able to compute optimized pricing policies without iteratively updating consumer behaviors under changing policies. Instead, the synchronized learning process allows to study the strategic interplay of competing vendors and various kinds of non-myopic consumer behavior.

Our reproducible experimental results demonstrate the applicability and the effectiveness of our models in different market scenarios. Further, the results allow to explain different phenomena and to analyze various performance comparisons. On the one hand, the results obtained provide insights into how to counteract different forms of strategic customer behavior by not losing too much rewards from myopic customer shares. On the other hand, we can verify and quantify the increase in consumer rent when using backward-looking reference price-based behaviors or forward-looking price-anticipating behaviors, respectively. The long-term performances of the policies obtained as well as associated price and sales trajectories over time are discussed for different market scenarios and managerial recommendations are inferred. Further, we identified limitations of the derived models and proposed promising model extensions and research directions for future work.

Table 7
Hyperparameters for Proximal Policy Optimization (PPO).

Parameter	Value
Learning rate	$3 \cdot 10^{-4}$
Steps per update	2048
Minibatch size	64
Epochs per update	10
clip_range (ϵ)	0.2
Discount factor (γ)	0.9999
Generalized advantage estimator factor (λ)	0.95
Entropy coefficient	0
Value function coefficient	0.5

CRediT authorship contribution statement

Fabian Lange: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Data curation, Conceptualization. **Rainer Schlosser:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

There are no conflicts of interest.

Appendix A

A.1. Hyperparameters

See Table 7

A.2. Notation table

See Table 8

Data availability

We provide an (anomymized) code repository for reproducibility:https://anonymous.4open.science/r/StrategicCustomerRL-6C1E.

Table 8
Overview of the notation of our main parameters and variables.

Symbol	Explanation
t	Current timestep
$n_{seasons}$	Number of seasons (cycle length)
$i_{season}(t)$	Index of season at time t, $i_{season}(t) \in \{0, \dots, n_{seasons} - 1\}$, $t = 0, 1, \dots$
$n_{vendors}$	Number of vendors
k	Vendor index, $k = 0, 1, \dots, n_{vendors}$ (0 for no buy option)
A	Admissible prices (action space)
a_t	Chosen action (of firm 1) at time t
$\vec{p}_t$	Vector of all price offers at time t, $\vec{p}_t = (p_t^{(1)}, \dots, p_t^{(n_{vendors})})$
$a^{comp}(p)$	Competitor offer price response to firm 1's current price p (duopoly case)
n_{last}	Length of considered price history in periods (parameter)
h	Index of time lag, $h = 1, \dots, n_{last}$
$\vec{p}_{t-h}$	Stored prices of all firms h timesteps ago (in time t), $h = 1, \dots, n_{last}$
$H_t^{(h)}$	Set of all stored prices for last h periods at time t, i.e., $H_t^{(h)} = (\vec{p}_{t-1}, \dots, \vec{p}_{t-h})$
n_c	Number of classes of (strategic) customer types
c	Customer types index, $c = 0, 1, \dots, n_c$ (0 for myopic)
$C(c)$	Average share of new arriving customers of type c, $c = 0, 1, \dots, n_c$ (parameter)
$N_t^{(c)}$	Number of customers eligible to buy per type c at t
$P_t^{(c,k)}(\vec{p}_t; H_t^{(h)})$	Probabilities of consumer type c to buy at firm k in t given $\vec{p}_t$ and history $H_t^{(h)}$
$w_t^{(c)}$	Number of waiting customers of type c at time t
$\vec{w}_t$	Vector of waiting customers at time t, $\vec{w}_t = (w_t^{(1)}, \dots, w_t^{(n_c)})$
i_{max}	Number of new arriving customer per period
$X_t^{(c)}$	Number of interested customers in period t of type c (random)
$Y_t^{(c,k)}$	Number of customers in period t of type c buying at vendor k (random)
$i_t^{(k)}$	Total number of customers buying at firm k in period t (random)
γ	Discount parameter (for on period)
r_t	Reward (of RL agent/firm 1) at time t
$\tilde{s}_t$	Full state of the system at time t, i.e., $\tilde{s}_t = (\vec{p}_t, H_t^{(max(n_{last}, \phi, \delta))}, \vec{w}_t)$
s_t	State of the system observable for firms at time t, i.e., $s_t = (\vec{p}_t, H_t^{(n_{last})})$
u_t	Utility score (myopic customer)
α, β_t	Scale parameter (myopic customer)
π_{remain}	Probability (per period) to stay in the waiting pool (type $c = 1, 2, 3$)
π_{return}	Probability (per period) to visit the market (type $c = 1, 2, 3$)
δ	Number of periods for past reference (price-aware customer)
h_{rel_ref}	Relative price threshold (price-aware customer)
WTP_{max}	Maximum willingness to pay (type $c = 2, 3$)
ϕ	Number of last periods for forecasting input (price-anticipating cust.)
θ	Forecasting horizon in periods (price-anticipating customer)
h_{rel_gap}	Relative price threshold (price-anticipating customer)
$AR(H_t^{(\phi)})$	Auto-regressive function for price forecasts given a price history $H_t^{(\phi)}$
φ_i	Parameter of $AR(H_t^{(\phi)})$, $i = 0, \dots, \phi$ (price-anticipating customer)
z_{t+i}	Price prediction at time t for timestep $t + i$, $i = 1/n_{vendors}, \dots, \theta$

References

[1] Micus C, et al. Methods to analyze customer usage data in a product decision process:A systematic literature review. Oper Res Perspect 2023;10:100277.
[2] Den Boer AV. Dynamic pricing and learning: Historical origins, current research, and new directions. Surv Oper Res Manag Sci 2015;20(1):1–18.
[3] McAfee RP, Te Velde V. Dynamic pricing in the airline industry. Handb Econ Inf Syst 2006;1:527–67.
[4] Abrate G, Fraquelli G, Viglia G. Dynamic pricing strategies: Evidence from European hotels. Int J Hosp Manag 2012;31(1):160–8.
[5] Assad S, et al. Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. J Political Econ 2024;132:723–71.
[6] Dutta G, Mitra K. A literature review on dynamic pricing of electricity. J Oper Res Soc 2017;68(10):1131–45.
[7] Chen K, et al. Dynamic pricing in the presence of reference price effect and consumer strategic behaviour. Int J Prod Res 2020;58(2):546–61.
[8] Lin Y-T, Parlaktürk AK, Swaminathan JM. Are strategic customers bad for a supply chain? Manuf Serv Oper Manag 2018;20:481–97.
[9] Shen M, Su X. Customer behavior modeling in revenue management and auctions: A review and new research opportunities. Prod Oper Manag 2007;16:713–28.
[10] Gönsch J, Klein R, Neugebauer M, Steinhardt C. Dynamic pricing with strategic customers. J Bus Econ 2012;83:505–49.
[11] Wei MM, Zhang F. Recent research developments of strategic consumer behavior in operations management. Comput Oper Res 2018;93:166–76.
[12] Kannan PK, Kopalle P. Dynamic pricing on the internet: Importance and implications for consumer behavior. Int J Electron Commer 2001;5:63–83.
[13] Haws K, Bearden W. Dynamic pricing and consumer fairness perceptions. J Consum Res 2006;33:304–11.
[14] Macgregor D. Augustin Cournot. The Mathematical Principles of the Theory of Wealth, 1838. Econ J 1838;39(153):91–2.
[15] Gerpott TJ, Berends J. Competitive pricing on online markets: A literature review. J Revenue Pricing Manag 2022;21(6):596–622.
[16] Bitran GR, Mondschein SV. Periodic pricing of seasonal products in retailing. Manag Sci 1997;43(1):64–79.
[17] Chatwin RE. Optimal dynamic pricing of perishable products with stochastic demand and a finite set of prices. European J Oper Res 2000;125(1):149–74.
[18] Dong J, Wu DD. Two-period pricing and quick response with strategic customers. Int J Prod Econ 2019;215:165–73, Emerging Issues in Multi-Channel Operations Management in the O2O Era.
[19] Granot D, Granot F, Mantin B. A dynamic pricing model under duopoly competition. Vancouver: Sauder School of Business, University of British Columbia; 2007.
[20] Kwon C, et al. Non-cooperative competition among revenue maximizing service providers with demand learning. European J Oper Res 2009;197(3):981–96.
[21] Gosavi A. Reinforcement learning: A tutorial survey and recent advances. INFORMS J Comput 2009;21(2):178–92.
[22] Pontrandolfo P, et al. Global supply chain management: A reinforcement learning approach. Int J Prod Res 2002;40(6):1299–317.
[23] Kephart JO, Hanson JE, Greenwald AR. Dynamic pricing by software agents. Comput Netw 2000;32(6):731–52.
[24] Gosavi A, Bandla N, Das TK. A reinforcement learning approach to a single leg airline revenue management problem with multiple fare classes and overbooking. IIE Trans 2002;34(9):729–42.
[25] Giannoccaro I, Pontrandolfo P. Inventory management in supply chains: a reinforcement learning approach. Int J Prod Econ 2002;78(2):153–61.
[26] Peters M, et al. A reinforcement learning approach to autonomous decision-making in smart electricity markets. Mach Learn 2013;92:5–39.
[27] Zhong S, et al. Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating. Appl Energy 2021;288:116623.
[28] Pandey V, Wang E, Boyles SD. Deep reinforcement learning algorithm for dynamic pricing of express lanes with multiple access locations. Transp Res Part C: Emerg Technol 2020;119:102715.
[29] Chen C, et al. Spatial-temporal pricing for ride-sourcing platform with reinforcement learning. Transp Res Part C: Emerg Technol 2021;130:103272.

[30] Lee S, Choi D-H. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach. Appl Energy 2021;304:117754.
[31] Liu D, et al. Dynamic pricing strategy of electric vehicle aggregators based on DDPG reinforcement learning algorithm. IEEE Access 2021;9:21556–66.
[32] Su X. Intertemporal pricing with strategic customer behavior. Manag Sci 2007;53(5):726–41.
[33] Coase RH. Durability and monopoly. J Law Econ 1972;15(1):143–9.
[34] Chen Y, Shi C. Joint pricing and inventory management with strategic customers. Oper Res 2019;67:1610–27.
[35] Zhang D, Cooper WL. Managing clearance sales in the presence of strategic customers. Prod Oper Manag 2008;17:416–31.
[36] Levina T, et al. Dynamic pricing with online learning and strategic consumers: An application of the aggregating algorithm. Oper Res 2009;57:327–41.
[37] Chen X, Gao J, Ge D, Wang Z. Bayesian dynamic learning and pricing with strategic customers. Prod Oper Manage 2022;31(8):3125–42.
[38] Famil Alamdar P, Seifi A. A deep Q-learning approach to optimize ordering and dynamic pricing decisions in the presence of strategic customers. Int J Prod Econ 2024;269:109154.
[39] Zhou Q, Yang Y, Fu S. Deep reinforcement learning approach for solving joint pricing and inventory problem with reference price effects. Expert Syst Appl 2022;195:116564.
[40] Zhou Q, et al. Joint pricing and inventory control with reference price effects and price thresholds: A deep reinforcement learning approach. Expert Syst Appl 2023;233:120993.
[41] Kastius A, Schlosser R. Dynamic pricing under competition using reinforcement learning. J Revenue Pricing Manag 2022;21:50–63.
[42] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: ICML 2018, 10-15, 2018. Proceedings of machine learning research, vol. 80, PMLR; 2018, p. 1856–65.
[43] Mnih V, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[44] Silver D, et al. Deterministic policy gradient algorithms. In: ICML'14, vol. I. 2014, p. 387–95.
[45] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods, CoRR abs/1802.09477. 2018.
[46] Mnih V, et al. Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. 2016, p. 1928–37.
[47] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, ArXiv Preprint arXiv:1707.06347.
[48] Shen Y, Zhang Q, Zhang Z, Ma X. Omnichannel retailing return operations with consumer disappointment aversion. Oper Res Perspect 2022;9:100253.
[49] Yadavalli VS, et al. An integrated optimization model for selection of sustainable suppliers based on customers' expectations. Oper Res Perspect 2019;6:100113.
[50] Buchanan JA. My reference point, not yours. J Econ Behav Organ 2020;171:297–311.
[51] Groeneveld J, et al. Self-learning agents for recommerce markets. Bus Inf Syst Eng 2024;66:441–63.