Mathematical Methods of Operations Research (2021) 94:1–34
https://doi.org/10.1007/s00186-021-00744-y

ORIGINAL ARTICLE

Dynamic pricing with finite price sets: a non-parametric approach

Athanassios N. Avramidis[1] · **Arnoud V. den Boer**[2,3]

Received: 29 April 2020 / Revised: 21 March 2021 / Accepted: 31 May 2021 / Published online: 28 June 2021

Abstract
We study price optimization of perishable inventory over multiple, consecutive selling seasons in the presence of demand uncertainty. Each selling season consists of a finite number of discrete time periods, and demand per time period is Bernoulli distributed with price-dependent parameter. The set of feasible prices is finite, and the expected demand corresponding to each price is unknown to the seller, whose objective is to maximize cumulative expected revenue. We propose an algorithm that estimates the unknown parameters in a learning phase, and in each subsequent season applies a policy determined as the solution to a sample dynamic program, which modifies the underlying dynamic program by replacing the unknown parameters by the estimate. Revenue performance is measured by the regret: the expected revenue loss relative to the optimal attainable revenue under full information. For a given number of seasons n, we show that if the number of seasons allocated to learning is asymptotic to $(n^2 \log n)^{1/3}$, then the regret is of the same order, uniformly over all unknown demand parameters. An extensive numerical study that compares our algorithm to six benchmarks adapted from the literature demonstrates the effectiveness of our approach.

Keywords Markov decision process · Dynamic programming · Dynamic pricing · Regret · Asymptotic analysis

Mathematics Subject Classification 60K10 · 93E35 · 90B05 · 62G20

✉ Athanassios N. Avramidis
aa1w07@soton.ac.uk

1 Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

2 Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

3 Amsterdam Business School, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands

1 Introduction

1.1 Background

Pricing of a perishable product is a central problem in many industries. As discussed by Talluri and van Ryzin (2005), a classical setting involves a firm facing successive seasons of finite length during which a finite fixed inventory is sold, and such that at the end of the season, unsold inventory expires worthless. The firm seeks to set prices in a way that maximizes the expected revenue. Instances of this problem are found in many industries, including fashion, retail, air travel, hospitality, and leisure. In Gallego and van Ryzin (1994), an optimal price is shown to be a function of the *state* (t, c) of the system, where t denotes remaining time to the end of the season and c denotes remaining inventory; this function increases with remaining time and decreases with remaining inventory. To compute these optimal prices, it is essential to know the relationship between price and expected demand—often referred to as the *demand function* or *demand curve*.

In practice, decision makers seldom have full knowledge about the demand process. The absence of full information about the demand process introduces a tension between demand learning ('exploration', by experimenting with selling prices) and revenue earning ('exploitation', i.e. using the estimated optimal prices). The longer one spends learning the demand properties, the less time remains to exploit that knowledge and earn revenue; on the other hand, less time spent on demand learning results in higher uncertainty that could diminish the revenue earned during the exploitation phase. A key feature of a good self-learning pricing algorithm is its ability to optimally balance this tension.

The problem of designing asymptotically optimal self-learning pricing algorithms has received considerable attention (see the literature review below). Several authors (e.g. Besbes and Zeevi 2009; Wang et al. 2014; Lei et al. 2014) have analyzed optimal pricing and learning with finite inventories in a particular asymptotic regime, where the performance of a price policy is evaluated when both the *expected demand per season and the initial inventory grow large*. In this so-called fluid regime, the problem is simplified by essentially removing the stochasticity of demand. This regime is well suited for applications where initial inventory and length of the selling season are large. However, in applications where initial inventory does not grow large, this asymptotic regime is not informative for a policy's performance. An informative example is ferry services. These are services that are regularly offered, with a finite selling season (tickets are sold until the departure of the ferry), finite inventory (determined by the size of the ferry), and with multiple selling seasons (corresponding to different days of departures). Another example comes from grocery retail: brick-and-mortar retail shops typically have a small inventory of each specific product to sell, and face many selling seasons during which a constant demand function might be postulated. Anecdotal evidence from the United Kingdom shows that it is not uncommon that the price is dropped several times, and, very near the closing time on the "use by" or "sell by" date, it is a small fraction of the original. In these examples, the relevant regime to study the performance of pricing algorithms is that of *repeated seasons with bounded*

inventory, and not a regime where the size of the ferry, or the food inventory, goes to infinity.

Perhaps closest to our work is the study by den Boer and Zwart (2015), who consider dynamic pricing with finite inventories in a setting with multiple, consecutive selling seasons, each with fixed, finite inventory. The authors assume a parametric demand model, characterized by two unknown parameters that are learned from accumulating sales data, and design and analyze an asymptotically near-optimal pricing algorithm. A disadvantage of their parametric approach is the risk of *model mis-specification*: large losses can be incurred if the true demand function is not of the assumed form (see Sect. 6 for a numerical illustration). To mitigate this risk a *non-parametric* approach is needed, that (i) does not restrict itself to a parametrized sub-class of demand functions, and (ii) performs well in a regime with consecutive selling seasons with *bounded* initial inventory. Such an approach is taken by this paper.

1.2 Overview of contributions

We consider a monopolist seller of a finite inventory of a perishable product, which is sold during consecutive selling seasons. In the same spirit as den Boer and Zwart (2015), we formulate a discrete-time, finite-state Markov Decision Process (MDP) in which the underlying state is the pair (inventory, time-to-perish); this MDP characterizes optimal pricing under knowledge of the demand function; it is the central element in our setting where the transition probabilities are unknown to the seller. We assume a finite number κ of feasible prices. In our basic model, each season has a length of T periods and the same initial inventory x—this assumption is later relaxed to allow for non-identical selling seasons. In any period during which the ith price is offered, the demand for the product is a Bernoulli(λ_i) random variable. The vector of demand rates (purchase probabilities) $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_\kappa) \in [0, 1]^\kappa$ is unknown to the seller. We emphasize that we make no assumptions on $\boldsymbol{\lambda}$.

The algorithm that we propose separates a given selling horizon of n seasons into an exploration and an exploitation phase. The exploration phase rotates the prices throughout, so that each price is applied (nearly) the same number of times (periods); it concludes with a (maximum likelihood) estimate $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\kappa)$ of $\boldsymbol{\lambda}$. A pricing policy is then constructed from the corresponding dynamic programming recursion, in which the unknown $\boldsymbol{\lambda}$ is replaced by the estimate $\widehat{\boldsymbol{\lambda}}$; we refer to this recursion as the *sample dynamic program*. The exploitation phase applies this policy throughout all the remaining time periods. It is noteworthy that this is not a fixed-price policy; instead, the price depends on the system state (t, c) which is constantly evolving. Our main result establishes that a carefully tuned length of the exploration phase implies that our policy is consistent, and has regret $O(n^2 \log n)^{1/3})$, uniformly over $\boldsymbol{\lambda}$ (see Theorem 1). Thus, the revenue generated by the proposed algorithm gets arbitrary close to the best achievable revenue under full knowledge of the unknown demand parameters, as n grows large. It is worth noting that this result holds without assuming that the initial inventory of some season grows large, as in antecedent literature. This theorem is then extended to a sequence of seasons that need not have the same initial

inventory or length (see Theorem 2); the extension merely requires that the sequences of inventory levels and season lengths are bounded.

We provide an extensive numerical study in which we compare the performance of our algorithm to six alternatives, based on four papers in the literature: algorithms based on the fluid approximation in Besbes and Zeevi (2012, Algorithm 1, Section 3.1); adaptations of the upper-confidence-bound approach of Babaioff et al. (2015); Thompson sampling (Ferreira et al. 2018, Algorithm 2); and the method of den Boer and Zwart (2015) adapted for a finite price set. For a wide variety of demand functions we show that our policy outperforms these alternatives; see Sect. 6 for more details.

1.3 Related literature

The literature on pricing strategies is vast. We refer to Bitran and Caldentey (2003); Elmaghraby and Keskinocak (2003); Talluri and van Ryzin (2005); Gallego and Topaloglu (2019) for comprehensive reviews on the subject. A recent survey and classification that focuses on pricing and learning appears in den Boer (2015).

This paper is related to literature that addresses demand learning in dynamic pricing problems. For a single-product setting without an inventory constraint, examples are Broder and Rusmevichientong (2012); den Boer and Zwart (2014); Besbes and Zeevi (2015), and Keskin and Zeevi (2014), who design and analyze self-learning pricing algorithms under a variety of demand models. Closer to this paper is a stream of literature in which inventory is finite; implying that the optimal price is not a single value but a function of the system state (remaining time and remaining inventory). In Lin (2006), Aviv and Pazgal (2005), Araman and Caldentey (2009), and Farias and Van Roy (2010), the demand function is characterized by a single unknown parameter that is learned in a Bayesian fashion. Besbes and Zeevi (2009); Wang et al. (2014); Lei et al. (2014) consider more general demand models, but assume an asymptotic regime where inventory grows large; as explained above, results derived in this regime are not informative for applications where inventory is bounded. Demand learning for non-perishable products is also related to our work. A recent example is Chen et al. (2019), who employ non-parametric demand learning towards joint pricing and inventory decisions. The no-perish assumption means that demand may be instantly met by an inventory unit that was procured at an arbitrary time point in the past, which makes these types of problems different from the one considered in this paper.

In the literature that addresses demand learning, a common approach is to estimate, based on accumulating sales data, the optimal solution to some fluid model (approximation of the stochastic control problem) efficiently enough to (nearly) minimize (revenue) losses asymptotically. One approach, exemplified by Besbes and Zeevi (2009, 2012), separates the selling season into disjoint pure-exploration (learning) and exploitation phases. A worst-case upper bound on losses is minimized by carefully selecting the amount of time to spend on learning, in a regime where the total expected demand and inventory level grow large at the same rate. A second approach formulates a multi-armed bandit problem, and deals with the exploration-exploitation tradeoff via a long-established method known as an *upper-confidence-bound* (Auer et al. 2002), and whose principle is "optimism in the face of uncertainty". Here, the usual maximum

likelihood estimates of demand are replaced by upper confidence bounds, and exploration and exploitation occur simultaneously. This approach is exemplified by Babaioff et al. (2015); Badanidiyuru et al. (2013). Babaioff et al. (2015) address the case of a continuous price set; their upper-confidence bounds apply to the expected revenue associated to each price, where the set of prices is asymptotically dense on the price domain. Badanidiyuru et al. (2013) address the case of a finite price set, and study a general model where rewards (revenue) and resource consumption are sampled from an unknown time-invariant distribution. Using upper- and lower-confidence bounds on mean rewards and mean resource consumptions, respectively, they aim to determine an optimal time-invariant *mix of prices*; optimality is with respect to the linear program (fluid approximation) in Besbes and Zeevi (2012). These papers characterize the regret through upper and lower bounds in a regime where expected demand grows to infinity. Ferreira et al. (2018) employ a randomized Bayesian method known as *Thompson sampling* whose aim is to learn efficiently a mix of prices that is optimal with respect to the same linear program. In a regime where mean demand grows to infinity, they upper-bound the *Bayesian regret*: the conditional average regret given a prior distribution on the demand vector.

The relationship between the current paper and this stream of literature can be summarized as follows: in analogy with this literature, in our model the aggregate amount of inventory and the aggregate mean demand over n seasons grow proportionally to n; however, this growth does not occur within one season of "large size", but instead through a *sequence of seasons* with bounded inventory and season length, and common demand function. This boundedness entails that even if one prices optimally with respect to some fluid model, one has not closed the *fluid gap*: the difference in expected revenue between the optimal policy and the fluid-optimal policy. This gap—which essentially arises by neglecting randomness of demand—may be negligible in cases where both inventory and length of the season are reasonably large. But, as observed by Maglaras (2011) (page 6),

> one would expect that the discrete and stochastic nature of the pricing problem to be [sic] relevant when selling 4 newly constructed single family homes over the course of 24 weeks, but it may be less relevant when selling 4000 pairs of skis over a similar time duration from, say, October to March.

Our model is designed precisely for settings where the fluid regime is not informative: that is, when inventory does not grow large but is finite (such as, e.g., in ferry services and grocery retail), and when neglecting the structure of the underlying MDP is detrimental in terms of revenue performance.

Perhaps closest to our work is den Boer and Zwart (2015), who study optimal pricing with multiple, consecutive selling seasons, each with finite initial inventory. In contrast to our paper, they work with a demand function of a known parametric form with two unknown parameters. They develop an (almost) certainty-equivalent strategy, which at all times maintains a parametric (quasi-maximum likelihood) estimate of the demand function, and prices optimally under the corresponding Markov decision process in which the unknown demand function is replaced by its estimate. The authors provide an upper bound on the regret after n selling seasons, and accompany this result by a lower bound that holds for any policy. A drawback of their parametric approach

is the risk of model mis-specification: in reality, demand may not be of the assumed parametric form, and pricing recommendations may be suboptimal. In our model we mitigate this drawback by making no assumption whatsoever on the shape of the demand function.

The remainder of this paper Section 2 formulates the problem, defines the regret, and discusses key differences with alternative approaches. Section 3 presents the proposed strategy, and Sect. 4 contains the asymptotic performance analysis. The extension to non-identical seasons appears in Sect. 5. Section 6 presents the results of our numerical study. A few auxiliary results appear in the Appendix.

Notation The notation ":=" stands for "is defined as". A statement such as $\lambda = \lambda(p) := f(p)$ means that the function $\lambda(p)$ is identical to $f(p)$, and that we may write λ instead of $\lambda(p)$ or $f(p)$. We use $\mathbb{N} := \{0, 1, 2, \ldots\}$ for the set of natural numbers. For a set A, A^c denotes the complement. Given any sample space Ω of which the generic element is denoted ω, and given a set $A \subseteq \Omega$, we write $\mathbb{1}_{[A]}$ for $\mathbb{1}_{[\omega \in A]}$, the random variable taking value 1 if the event A occurs, and 0 otherwise. For any real z, we write $\lfloor z \rfloor$ for the *floor*, the largest integer that is no larger than z; $\lceil z \rceil$ for the *ceiling*; $[z]$ for the integer nearest to z; and z^+ for the *positive part* $\max\{0, z\}$. For any vector $\mathbf{z} = (z_i)$ we define $\|\mathbf{z}\| := \max_i |z_i|$. A sum over an empty index set, for example $\sum_{i=1}^{0} z_i$, is understood as zero. With a_n and b_n being nonnegative sequences, we write $a_n = O(b_n)$ if a_n/b_n is bounded from above by a constant; we write $a_n = \Omega(b_n)$ if a_n/b_n is bounded from below by a constant; and if a_n/b_n is bounded from both above and below, then we write $a_n \asymp b_n$.

2 Problem formulation

Basic elements We consider a monopolist seller of perishable products which are sold during consecutive selling seasons. Each season has a positive integer *length* of T (indivisible) time periods. At the start of each selling season the seller has a positive integer *inventory* (inventory) of x units, which can only be sold during that particular season. At the end of each season, any unsold inventory is worthless, and its disposal costs nothing. In our basic model, identical such seasons occur in succession: the ith season consists of the time periods indexed from $(i-1)T+1$ to iT, for all $i = 1, 2, \ldots, n$, where n is a selling horizon that is known at time zero. In Sect. 5 we relax this assumption, and consider non-identical seasons.

There are κ distinct actions corresponding to prices that increase in the action index: $0 < p_1 < p_2 < \ldots < p_\kappa < \infty$. There is additionally a *shutoff* action, indexed 0, whose sole function is to shut off the demand; the price associated to this action is immaterial (since no sale is ever made); thus we set $p_0 = 0$ without loss of generality. In each period u, the seller chooses an *action* $A_u \in \mathcal{A}$, where $\mathcal{A} = \{0, 1, 2, \ldots, \kappa\}$, and thus sets the price to p_{A_u}. After setting the price, a binary demand is observed, which indicates whether one unit is sold or not.

The demand is stochastic and price-dependent. Write D_u for the demand in period u, and define the set $\mathcal{H}_u := \{(a_1, \ldots, a_u, d_1, \ldots, d_u) \in \mathcal{A}^u \times \{0, 1\}^u\}$, for all $u \in \mathbb{N}$, and $\mathcal{H}_0 := \emptyset$. For each $u = 1, \ldots, nT$, each element $(a_1, \ldots, a_u, d_1, \ldots, d_u)$ of $\mathcal{H}_u$ is a potential *history* of prices and demand that the seller might observe in the

first u time periods. A *(pricing) strategy* $\sigma = (\sigma_u)_{u\in\mathbb{N}}$ is a collection of functions $\sigma_u : \mathcal{H}_{u-1} \to \mathcal{A}$ such that at each time $u = 1, \ldots, nT$, the seller's action is $A_u = \sigma_u(A_1, D_1, \ldots, A_{u-1}, D_{u-1})$. Thus, the policy specifies, for each possible data set of previously used prices and corresponding demand observations, which price should be used in the next time period.

Our main assumption with respect to the demand mechanism is that each action (price) $a \in \mathcal{A}$ is associated to a probability of purchase, λ_a, which is unknown to the seller, except for the shut-off property $\lambda_0 = 0$. Specifically, we assume that, conditionally on $A_u = a$, D_u is Bernoulli distributed with mean λ_a, for all $a \in \mathcal{A}$, and is independent of past actions and demands $\{A_1, \ldots, A_{u-1}, D_1, \ldots, D_{u-1}\}$. The vector of purchase probabilities, $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_\kappa)$, is unknown to the seller. We write $\boldsymbol{\Lambda} := [0, 1]^\kappa$ for the set of all possible purchase probability vectors.

To describe the dynamics of the seller's remaining inventory, observe that any period u is contained in the season numbered $\lceil u/T \rceil$ and corresponds to the *seasonal remaining time* $t_u := \lceil u/T \rceil T - u + 1 \in \{1, \ldots, T\}$, which is the number of periods that remain in the season containing period u. For example, for $T = 10$, the period indexed $u = 11$ is contained in season $\lceil u/T \rceil = 2$ and corresponds to a seasonal remaining time $t_{11} = 2 \cdot 10 - 11 + 1 = 10$. The end of any season coincides with the beginning of a new season; at any such boundary, any unused inventory from the ending season expires worthless and at no cost; the inventory of the new season is replenished to x, and the seasonal remaining time t_u becomes T. The *inventory* at the beginning of period u is denoted C_u throughout; in the basic model, it evolves as follows:

$$C_u = \left\{ \begin{array}{ll} x & \text{if } t_u = T \\ \max\{C_{u-1} - D_{u-1}, 0\} & \text{otherwise} \end{array} \right\}, \quad u = 1, 2, \ldots \tag{1}$$

The revenue earned in any period u is $p_{A_u} \min\{C_u, D_u\} = p_{A_u} \mathbb{1}_{[C_u>0]} D_u$. Given a planning horizon of n seasons, the seller's objective is to determine a strategy σ that maximizes the expected revenue, $\sum_{u=1}^{nT} \mathbb{E}_\sigma[p_{A_u} \min\{C_u, D_u\}]$, where $\mathbb{E}_\sigma[\cdot]$ denotes the expectation under strategy σ.

Optimal solution under full information Provided the probability vector $\boldsymbol{\lambda}$ is known, an optimal pricing strategy can be determined by solving a Markov Decision Process (MDP) corresponding to a *single* selling season. The states, transitions, and rewards of this MDP are defined as follows. A *state* (t, c) encodes that the seasonal remaining time is t and the remaining inventory is c. The set of states is $\mathcal{X} = \{(t, c) : t \in \{0, 1, \ldots, T\}, c \in \{0, 1, \ldots, x\}\}$. The transition dynamics depend on the actions taken, and are as follows. If action a is used in state (t, c), then with probability λ_a a state transition $(t, c) \to (t - 1, (c - 1)^+)$ occurs, and revenue $p_a \mathbb{1}_{[c>0]}$ is earned; with probability $1 - \lambda_a$ a state transition $(t, c) \to (t - 1, c)$ occurs, and no revenue is earned.

A *policy* π is a set of actions at all states: $\pi = (\pi_{t,c})_{(t,c)\in\mathcal{X}}$ with $\pi_{t,c} \in \mathcal{A}$ for each $(t, c) \in \mathcal{X}$. The set of all policies is denoted Π, and is finite. Given any policy π and state $(t, c) \in \mathcal{X}$, the *value* of the state is the expected revenue-to-go (to the end of the season) when starting in this state and using the actions of π; it is denoted $V^\pi_{t,c}$. These values satisfy the recursion

$$\left\{ \begin{array}{l} V^{\pi}_{t,c} = (1-\lambda_{\pi_{t,c}})V^{\pi}_{t-1,c} + \lambda_{\pi_{t,c}}[p_{\pi_{t,c}} + V^{\pi}_{t-1,c-1}] \\ \quad\quad = \lambda_{\pi_{t,c}}[p_{\pi_{t,c}} - \Delta V^{\pi}_{t-1,c}] + V^{\pi}_{t-1,c} \end{array} \right\} \quad 1 \le c \le x, \quad t = 1, 2, \ldots, T, \tag{2}$$

where $\Delta V^{\pi}_{t-1,c} := V^{\pi}_{t-1,c} - V^{\pi}_{t-1,c-1}$ for $c \ge 1$; $V^{\pi}_{t,0} := 0$ for all t; and $\Delta V^{\pi}_{0,c} := V^{\pi}_{0,c} = 0$ for all c.

By the finiteness of Π, there exists an *optimal policy* $\pi^* \in \Pi$ that maximizes $V^{\pi}_{T,x}$. The *optimal value* at a state (t, c) is the maximum expected revenue-to-go, starting from that state; it is denoted $V_{t,c}$. The values V and the policy π^* are determined recursively, backward in time:

$$\left\{ \begin{array}{l} V_{t,c} = \max_{a\in\mathcal{A}} \lambda_a[p_a - \Delta V_{t-1,c}] + V_{t-1,c} \\ \pi^*_{t,c} = \min \arg\max_{a\in\mathcal{A}} \lambda_a[p_a - \Delta V_{t-1,c}] \end{array} \right\} \quad 1 \le c \le x, \quad t = 1, 2, \ldots, T, \tag{3}$$

where $\Delta V_{t,c} := V_{t,c} - V_{t,c-1}$ for $c \ge 1$, $V_{t,0} = 0$ for all t; and $\Delta V_{0,c} = V_{0,c} = 0$ for all c. The number $V_{T,x} := V_{T,x}(\boldsymbol{\lambda})$ is the maximum possible expected revenue of a seller that knows $\boldsymbol{\lambda}$, for a season of length T and inventory x. By the (conditional) independence of demand across seasons, an optimal strategy consists of applying π^* in each season $s = 1, \ldots, n$.

Performance measure The *regret* of a strategy σ over the first n selling seasons is defined as $\mathcal{R}_n := \mathcal{R}_n(\sigma; \boldsymbol{\lambda}) := \mathcal{R}_n(\sigma, \boldsymbol{\lambda}, x, T) := nV_{T,x} - \sum_{u=1}^{nT} \mathbb{E}_{\sigma}[p_{A_u} \min\{C_u, D_u\}]$; it depends on the unknown $\boldsymbol{\lambda}$, and also on x and T. The regret is the (expected) revenue loss incurred by strategy σ relative to the optimal strategy of using the policy π^* in each season. The regret is based on an integer number of seasons, rather than an integer number of periods; this is natural, since policy (and revenue) optimality is with respect to a whole season and not any individual period. In our numerical study we mainly work with the *relative regret*, defined and denoted as $\mathcal{R}'_n := \mathcal{R}'_n(\sigma) := \mathcal{R}'_n(\sigma; \boldsymbol{\lambda}, x, T) := \mathcal{R}_n(\sigma, \boldsymbol{\lambda}, x, T)/(nV_{T,x})$. By definition, the value of the relative regret is a number that always lies in the interval [0, 1]; the smaller its value, the better the performance of σ; a value of zero indicates that σ extracts the maximum possible revenue.

3 Proposed pricing strategy

In this section we propose a data-driven pricing strategy that learns the optimal policy defined in Sect. 2. The strategy divides the time horizon into two phases, an exploration phase and an exploitation phase. In the exploration phase, all prices are used a nearly equal number of times, and the obtained sales data is used to construct an estimate $\widehat{\boldsymbol{\lambda}}$ of the unknown demand vector $\boldsymbol{\lambda}$. In the exploitation phase, the policy π^* defined in Sect. 2, with $\boldsymbol{\lambda}$ replaced by its estimate $\widehat{\boldsymbol{\lambda}}$, is used in all remaining selling seasons.

More specifically, given an estimate $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\kappa)$ (purely from the exploration phase here; this is relaxed later), the policy that is used throughout the exploitation phase is the solution of the sample dynamic program:

$$\left\{\begin{array}{l}\widehat{V}_{t,c} = \max_{a\in\mathcal{A}} \widehat{\lambda}_a[p_a - \Delta\widehat{V}_{t-1,c}] + \widehat{V}_{t-1,c},\\ \hat{\pi}_{t,c} = \min \arg\max_{1\le a\le\kappa} \widehat{\lambda}_a[p_a - \Delta\widehat{V}_{t-1,c}]\end{array}\right\} \quad 1 \le c \le x, \quad t = 1, 2, \ldots, T, \tag{4}$$

where $\Delta\widehat{V}_{t,c} := \widehat{V}_{t,c} - \widehat{V}_{t,c-1}$, $\widehat{V}_{t,0} := 0$ for all t, and $\Delta\widehat{V}_{0,c} := \widehat{V}_{0,c} = 0$ for all c. In particular, the shutoff action is excluded at all states with $t \ge 1$ and $c \ge 1$. We denote this policy as $\hat{\pi}$, or, to make the dependence on $\widehat{\boldsymbol{\lambda}}$ explicit, as $\pi(\widehat{\boldsymbol{\lambda}})$.

For $i \in \{1, \ldots, \kappa\}$ and $\tau \in \mathbb{N}$, the *(price-specific) sample size* is defined as $N_i(\tau) := \sum_{u=1}^{\tau} \mathbb{1}_{[A_u=i]}$; it is the count of time periods up to (including) τ during which the price is p_i.

STRATEGY $\sigma(\tau)$.

Step 1 (Initialization). Let $\tau \in \mathbb{N}$, $\tau \le n$.
Step 2 (Exploration).

(a) For all $u = 1, \ldots, \tau T$: if $C_u > 0$, then set A_u as the action i for which $N_i(u-1)$ is the smallest (in case of a tie, select the price with the lowest index); formally, $A_u := \min\{\arg\min\{N_i(u-1) : 1 \le i \le \kappa\}\}$. If $C_u = 0$, set $A_u = 0$.
(b) For each $i = 1, \ldots, \kappa$, let $N_i := N_i(\tau T) = \sum_{u=1}^{\tau T} \mathbb{1}_{[A_u=i]}$ be the count of time periods in the first τ seasons during which the price was p_i, and let $S_i := \sum_{u=1}^{\tau T} \mathbb{1}_{[A_u=i]} D_u$ be the count of sales obtained in these periods. Set $\widehat{\lambda}_i := S_i N_i^{-1} \mathbb{1}_{[N_i>0]}$, $i = 1, \ldots, \kappa$, and $\widehat{\boldsymbol{\lambda}} := (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\kappa)$.

Step 3 (Exploitation). For each season $s = \tau + 1, \ldots, n$, apply the policy $\pi(\widehat{\boldsymbol{\lambda}})$ defined in (4).

Step 2(a) ensures a near-parity of price-specific sample sizes at all times (the motivation for this is seen in proofs that follow). On a high level, this strategy is reminiscent of classical explore-then-commit policies of multi-armed bandit problems; see, e.g., Lattimore and Szepesvári (2019), Chapter 6. These policies divide the time horizon into two phases. In the first phase all actions are tried a number of times, in order to estimate the expected revenue associated to each action. In the second phase, an action with the highest estimated expected revenue is used at all times. Our strategy loosely adapts this idea to the MDP in Sect. 2: an optimal 'action' (of the multi-armed bandit problem) corresponds to an optimal policy for the MDP here. In the exploitation phase, we thus use the estimated state-dependent optimal prices (i.e., the estimated optimal policy $\hat{\pi}$) and not a fixed price at all times.

4 Performance analysis

4.1 Upper bound

In this section we show that the prices generated by our pricing strategy converge to the optimal prices corresponding to the MDP defined in Sect. 2, as the number of selling seasons n grows large. More precisely, we prove that the regret of strategy $\sigma(\tau_n)$ is bounded above by a constant times $n^{2/3}\log(n)^{1/3}$, under a suitable choice of the

exploration length τ_n. The constant depends only on x and T and grows at most linearly in each. This bound holds uniformly over all probability vectors $\boldsymbol{\lambda}$. Equivalently, the relative regret converges to zero at rate $O((\log(n)/n)^{1/3})$, uniformly over all $\boldsymbol{\lambda}$.

Theorem 1 *Set $\tau_n \asymp (n^2 \log n)^{1/3}$. Then, there exists a finite positive constant K_1 such that, for all $n \geq 2$,*

$$\sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \mathcal{R}_n(\sigma(\tau_n); \boldsymbol{\lambda}) \leq K_1 (n^2 \log n)^{1/3}.$$

To prove the theorem, we first provide a bound on the estimation error (Proposition 1) and a bound on the effect of this error during the exploitation phase (Proposition 2). Let $\widehat{\boldsymbol{\lambda}}_n$ be the estimator of $\boldsymbol{\lambda}$ corresponding to $\sigma(\tau_n)$; that is, obtained after an exploration period consisting of τ_n seasons. Recall that $N_i(\tau_n T) = \sum_{u=1}^{\tau_n T} \mathbb{1}_{[A_u = i]}$, for $i = 1, 2, \ldots, \kappa$, is the number of times up to the end of the learning phase that the price on offer is p_i.

Proposition 1 (Estimation error) *Let $f := \min\{x, T\}/\kappa$. For any $n \in \mathbb{N}$ and $\delta > 0$, we have*

$$\mathbb{E} \left\| \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda} \right\| \leq \delta + 2\kappa \exp(-2 \lfloor f \tau_n \rfloor \delta^2). \tag{5}$$

Proof of Proposition 1 Let $n \in \mathbb{N}$ and $\delta > 0$. We first obtain a lower bound on $N_i = N_i(\tau_n T)$ for each i. Let u denote a period of the learning phase such that the inventory is positive, that is, $1 \leq u \leq \tau_n T$ and $C_u > 0$. Any such period u contributes one unit to the sum $\sum_{i=1}^{\kappa} N_i$; that is, $\sum_{i=1}^{\kappa} N_i(u) = 1 + \sum_{i=1}^{\kappa} N_i(u-1)$. We claim that

$$\sum_{i=1}^{\kappa} N_i = \sum_{i=1}^{\kappa} N_i(\tau_n T) \overset{(a)}{\geq} \tau_n \min\{T, x\}, \quad \text{and} \quad |N_i - N_j| \overset{(b)}{\leq} 1 \text{ for } i \neq j. \tag{6}$$

Inequality (a) holds because the learning phase consists of τ_n seasons, in each of which there are at least $\min\{T, x\}$ periods u such that $C_u > 0$ (since there are x units initially, T sale periods, and no more than one unit is sold per period). Inequality (b) is the near-parity of sample sizes across prices, which is ensured by step 2(a) in the definition of σ. Now (6) implies

$$N_i \geq \lfloor f \tau_n \rfloor \quad \text{for each } i = 1, \ldots, \kappa. \tag{7}$$

Now define the event $\mathcal{E}_n := \{\left\| \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda} \right\| \leq \delta\}$. We have

$$\begin{aligned} \mathbb{E} \left\| \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda} \right\| &= \mathbb{E}[\left\| \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda} \right\| | \mathcal{E}_n^c] \mathbb{P}(\mathcal{E}_n^c) + \mathbb{E}[\left\| \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda} \right\| | \circ \mathcal{E}_n^c] \mathbb{P}(\circ \mathcal{E}_n^c) \\ &\leq \delta + \mathbb{P}(\circ \mathcal{E}_n^c) \\ &\overset{(a)}{\leq} \delta + 2\kappa \exp(-2 \lfloor f \tau_n \rfloor \delta^2), \end{aligned} \tag{8}$$

where step (a) is justified as follows:

$$\begin{aligned}\mathbb{P}(\circ\mathcal{E}_n^c) &= \mathbb{P}\big(\cup_{i=1}^{\kappa}\{|\widehat{\lambda}_{n,i}-\lambda_i|>\delta\}\big)\\ &\overset{(b)}{\le}\sum_{i=1}^{\kappa}\mathbb{P}(|\widehat{\lambda}_{n,i}-\lambda_i|>\delta)\overset{(c)}{\le}2\kappa\exp(-2\lfloor f\tau_n\rfloor\delta^2),\end{aligned}\tag{9}$$

where step (b) follows from a union bound. To justify step (c), observe that $\widehat{\lambda}_{n,i}$ is the sample mean of $N_i \ge \lfloor f\tau_n\rfloor$ i.i.d. Bernoulli(λ_i) random variables, and by Hoeffding's inequality, if $\{I_i\}_{i=1}^m$ are independent Bernoulli(q) random variables with $q \in (0,1)$, then, for any $m \ge 1$ and $\delta > 0$ we have $\max\big\{\mathbb{P}\big(\sum_{i=1}^m I_i - mq \ge m\delta\big), \mathbb{P}\big(\sum_{i=1}^m I_i - mq \le -m\delta\big)\big\} \le e^{-2m\delta^2}$. □

Next, we bound the loss incurred by policy $\pi(\widehat{\boldsymbol{\lambda}})$ against the optimal one.

Proposition 2 (Effect of estimation error) *Let* $\overline{p} := \max_{a\in\mathcal{A}} p_a$. *Then*

$$\max_c(V_{t,c}-V_{t,c}^{\hat{\pi}}) \le 4\overline{p}t\max_a|\widehat{\lambda}_a-\lambda_a| \quad \textit{for all } t=1,2,\ldots,T.\tag{10}$$

Proof of Proposition 2 Let $\epsilon := \max_a|\widehat{\lambda}_a-\lambda_a|$. We will prove two results: the value estimates $\widehat{V}$ are close to the optimal values:

$$V_{t,c}-\widehat{V}_{t,c} \le 2\epsilon t\overline{p} \quad \text{for all } t=1,2,\ldots,T \text{ and all } c=0,1,\ldots,x;\tag{11}$$

and the values $V^{\hat{\pi}}$ of the policy associated to $\widehat{V}$ are close to these estimates:

$$\widehat{V}_{t,c}-V_{t,c}^{\hat{\pi}} \le 2\epsilon t\overline{p} \quad \text{for all } t=1,2,\ldots,T \text{ and all } c=0,1,\ldots,x.\tag{12}$$

For all (t,c) such that $t=0$ or $c=0$, $V_{t,c}=\hat{V}_{t,c}$. In addition, for all c,

$$V_{1,c}=\max_a\lambda_a p_a \le \max_a(\hat{\lambda}_a+\epsilon)p_a \le \hat{V}_{1,c}+\epsilon\overline{p},$$

so that $V_{1,c}-\hat{V}_{1,c}\le 2\epsilon\overline{p}$. Now let $t\ge 1$ and suppose that $V_{t,c}-\hat{V}_{t,c}\le 2\epsilon t\overline{p}$, for all c. Then, for all actions a,

$$\begin{aligned}&\lambda_a(p_a+V_{t,c-1})+(1-\lambda_a)V_{t,c}\\ &\quad=\hat{\lambda}_a(p_a+\hat{V}_{t,c-1})+(1-\hat{\lambda}_a)\hat{V}_{t,c}\\ &\qquad+(\lambda_a-\hat{\lambda}_a)p_a+(1-\hat{\lambda}_a)(V_{t,c}-\hat{V}_{t,c})+\hat{\lambda}_a(V_{t,c-1}-\hat{V}_{t,c-1})\\ &\qquad+(\hat{\lambda}_a-\lambda_a)(V_{t,c}-V_{t,c-1})\\ &\quad\le\hat{\lambda}_a(p_a+\hat{V}_{t,c-1})+(1-\hat{\lambda}_a)\hat{V}_{t,c}+\epsilon p_a+(1-\hat{\lambda}_a)\cdot(2\epsilon t\overline{p})+\hat{\lambda}_a\cdot(2\epsilon t\overline{p})+\epsilon\cdot\overline{p}\\ &\quad\le\hat{\lambda}_a(p_a+\hat{V}_{t,c-1})+(1-\hat{\lambda}_a)\hat{V}_{t,c}+2\epsilon(t+1)\overline{p},\end{aligned}$$

using $|V_{t,c} - V_{t,c-1}| \le \overline{p}$, so that for all $c \ge 1$,

$$\begin{aligned} V_{t+1,c} &= \max_a \lambda_a(p_a + V_{t,c-1}) + (1-\lambda_a)V_{t,c} \\ &\le \max_a \hat{\lambda}_a(p_a + \hat{V}_{t,c-1}) + (1-\hat{\lambda}_a)\hat{V}_{t,c} + 2\epsilon(t+1)\overline{p} = \hat{V}_{t+1,c} + 2\epsilon(t+1)\overline{p}. \end{aligned}$$

This proves (11).

We now consider (12). For all (t,c) such that $t = 0$ or $c = 0$, $V^{\hat{\pi}}_{t,c} = \widehat{V}_{t,c} = 0$. Now suppose for induction that $\widehat{V}_{t,c} - V^{\hat{\pi}}_{t,c} \le 2\epsilon t\overline{p}$ for all $c \ge 0$. Then for all $c \ge 1$,

$$\begin{aligned} \hat{V}_{t+1,c} &= \hat{\lambda}_{\hat{\pi}_{t+1,c}}(p_{\hat{\pi}_{t+1,c}} + \hat{V}_{t,c-1}) + (1-\hat{\lambda}_{\hat{\pi}_{t+1,c}})\hat{V}_{t,c} \\ &= \lambda_{\hat{\pi}_{t+1,c}}(p_{\hat{\pi}_{t+1,c}} + V^{\hat{\pi}}_{t,c-1}) + (1-\lambda_{\hat{\pi}_{t+1,c}})V^{\hat{\pi}}_{t,c} \\ &\quad + \lambda_{\hat{\pi}_{t+1,c}}(\hat{V}_{t,c-1} - V^{\hat{\pi}}_{t,c-1}) + (1-\lambda_{\hat{\pi}_{t+1,c}})(\hat{V}_{t,c} - V^{\hat{\pi}}_{t,c}) \\ &\quad + (\hat{\lambda}_{\hat{\pi}_{t+1,c}} - \lambda_{\hat{\pi}_{t+1,c}}) \cdot (p_{\hat{\pi}_{t+1,c}} + \hat{V}_{t,c-1} - \hat{V}_{t,c}) \\ &\le \lambda_{\hat{\pi}_{t+1,c}}(p_{\hat{\pi}_{t+1,c}} + V^{\hat{\pi}}_{t,c-1}) + (1-\lambda_{\hat{\pi}_{t+1,c}})V^{\hat{\pi}}_{t,c} \\ &\quad + \lambda_{\hat{\pi}_{t+1,c}} \cdot 2\epsilon t\overline{p} + (1-\lambda_{\hat{\pi}_{t+1,c}}) \cdot 2\epsilon t\overline{p} + \epsilon \cdot (2\overline{p}) \\ &= V^{\hat{\pi}}_{t,c} + 2\epsilon(t+1)\overline{p}, \end{aligned}$$

using $|\hat{V}_{t,c-1} - \hat{V}_{t,c}| \le \overline{p}$. This proves (12). Now (10) follows directly from (11) and (12). □

We now prove Theorem 1 in two steps. First, we apply Propositions 1 and Proposition 2 to obtain an upper bound on the regret incurred during the exploitation phase. The regret incurred during the exploration phase is upper-bounded by a constant times the duration of this phase. Then, we show that the choice $\tau_n \asymp (n^2 \log n)^{1/3}$ minimizes the order of this upper bound, and obtain the $O(n^2 \log n)^{1/3})$ upper bound on the regret.

Proof of Theorem 1 Let $n \in \mathbb{N}$, $n \ge 2$. We proceed in two steps.

Step 1. Let $\widehat{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\lambda}}_n = (\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,\kappa})$ be the estimate obtained in the exploration phase of $\sigma(\tau_n)$. Let $V := V_{T,x}$ and $V^{\hat{\pi}}(\widehat{\boldsymbol{\lambda}}_n) := V^{\hat{\pi}}_{T,x}(\widehat{\boldsymbol{\lambda}}_n)$. Then

$$\begin{aligned} \mathcal{R}_n(\sigma(\tau_n); \boldsymbol{\lambda}) &\overset{(a)}{\le} \tau_n V + \mathbb{E}\big[(n-\tau_n)\big(V - V^{\hat{\pi}}(\widehat{\boldsymbol{\lambda}}_n)\big)\big] \\ &\overset{(b)}{\le} \tau_n V + n \cdot 4T\overline{p} \cdot \mathbb{E}\left\|\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}\right\|. \end{aligned} \tag{13}$$

Here (a) is argued as follows: the learning phase consists of τ_n seasons, in each of which the expected loss relative to the optimum is at most V. The exploitation phase consists of $n - \tau_n$ seasons, in each of which the conditional expected loss relative to the optimum, given $\widehat{\boldsymbol{\lambda}}_n$, is $V - V^{\hat{\pi}}(\widehat{\boldsymbol{\lambda}}_n)$. Step (b) then follows directly from Proposition 2.

Step 2. From (13) and Proposition 1 we obtain

$$\mathcal{R}_n \le \tau_n V + n \cdot 4T\overline{p} \cdot \big[\delta + 2\kappa \exp(-2\lfloor f\tau_n \rfloor \delta^2)\big], \tag{14}$$

for any n and $\delta > 0$. The assumption $\tau_n \asymp (n^2 \log n)^{1/3}$ implies that there are positive constants $\underline{c}_\tau$ and $\overline{c}_\tau$, such that, for all $n \geq 1$,

$$\underline{c}_\tau (n^2 \log n)^{1/3} \leq \tau_n \leq \overline{c}_\tau (n^2 \log n)^{1/3}. \tag{15}$$

Each of the summands in the right side of (14) is $O(n^2 \log n)^{1/3}$, provided that

$$\delta = \delta_n := c_\delta \left(\frac{\log n}{n} \right)^{1/3}, \quad \text{where } c_\delta = (6 f \underline{c}_\tau)^{-1/2}. \tag{16}$$

For the term $\tau_n V$ in (14), this follows directly from (15). For the second term on the right side of (14), note that $n \cdot 4T\overline{p} \cdot \delta_n \leq 4T\overline{p} \cdot c_\delta \cdot (n^2 \log n)^{1/3}$, and

$$e^{-2\lfloor f\tau_n \rfloor \delta_n^2} \leq e^{-2(f\tau_n - 1)\delta_n^2} \leq K_0 e^{-2f\tau_n \delta_n^2} \overset{(a)}{\leq} K_0 e^{-2f\underline{c}_\tau c_\delta^2 \log n} \overset{(b)}{=} K_0 n^{-1/3}; \tag{17}$$

here $K_0 := \sup_{n \geq 2} \exp(2\delta_n^2) = \exp(2c_\delta^2 (\log(3)/3)^{2/3})$, step (a) follows from the lower bound in (15) combined with (16) (since $\tau_n \delta_n^2 \geq \underline{c}_\tau c_\delta^2 \log n$), and step (b) follows from the definition of c_δ. Putting all terms together, we obtain $\sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \mathcal{R}_n(\sigma(\tau_n); \boldsymbol{\lambda}) \leq K_1 (n^2 \log n)^{1/3}$, where

$$K_1 := \sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} V(\boldsymbol{\lambda}) \overline{c}_\tau + 4T\overline{p} \big[(6 f \underline{c}_\tau)^{-1/2} + 2\kappa K_0 (\log 2)^{-1/3} \big].$$

But $K_1 < \infty$, since $\sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} V(\boldsymbol{\lambda}) \leq \overline{p} \min\{x, T\}$. □

Remark 1 A choice of τ_n that is consistent with Theorem 1 is

$$\tau_n = \lceil c_\tau (n^2 \log n)^{1/3} \rceil, \quad \text{where } c_\tau = \frac{1}{2} (3f)^{-1/3}. \tag{18}$$

To motivate the formula, observe that (17) shows that the exponential term $\exp\big(-2\lfloor f\tau_n \rfloor \delta_n^2\big)$ in (14) is $O(n^{-1/3})$, and therefore

$$\limsup_{n \to \infty} \frac{\sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \mathcal{R}_n(\sigma(\tau_n); \boldsymbol{\lambda})}{(n^2 \log n)^{1/3}} \leq \overline{p} \max\{\min\{x, T\}, 4T\} \big[\overline{c}_\tau + (6 f \underline{c}_\tau)^{-1/2} \big]. \tag{19}$$

The right side is minimized by setting $\underline{c}_\tau = \overline{c}_\tau$ and minimizing with respect to this single variable; this yields the value in (18).

4.2 Strength of bound

If $T = 1$, our problem reduces to a conventional multi-armed bandit problem. It is known (see, e.g., Lattimore and Szepesvári (2019), Exercise 15.6) that in this setting, the (worst-case) regret of explore-then-commit type of strategies grows as $n^{2/3}$. This implies that the $n^{2/3}$ growth rate (up to logarithmic terms) in Theorem 1 cannot be

improved by more refined proof techniques, but is an intrinsic property of the strategy σ.

It is also known that in multi-armed bandit problems with $K \in \mathbb{N}$ actions (arms), strategies such as MOSS Audibert and Bubeck (2009) achieve $O(\sqrt{Kn})$ worst-case regret, and this rate is the best possible (Vogel 1960; Auer et al. 2002). Neither this policy nor this characterization of the best possible growth rate of regret are directly transferable to an informative statement in our setting: if we would naively treat our problem as a multi-armed bandit problem, then each of the K arms in the multi-armed bandit problem would correspond to a policy π as defined in Sect. 2; as a result, the number of actions would be $K = \kappa^{T \cdot x}$ and hence the lower bound $\sqrt{Kn}$ would be prohibitively large in many instances that are practically relevant. For example, in our numerical study in Sect. 6 we consider $\kappa = 10$, $x = 100$, and $T = 65$, which could correspond to 10^{6500} different policies. There do exist algorithms for multi-armed bandit problems with an underlying MDP structure (e.g. Burnetas and Katehakis 1997; Even-Dar et al. 2006; Auer and Ortner 2007). Specific to our problem is that the transition probabilities of the MDP are unknown and governed by the same unknown parameters $\boldsymbol{\lambda}$, for each state (where inventory is available); this particular structure is exploited by the design of σ. Furthermore, Even-Dar et al. (2006, Theorem 13) provides upper and lower bounds (holding with high probability) on the value functions of a finite-state MDP; these bounds grow linearly in the time horizon, matching the growth rate of the multiplier $(2\overline{p}t)$ that we establish in Proposition 2. This suggests that the linear dependence of regret on the time horizon cannot be improved.

It is also insightful to compare our regret bound of Theorem 1 to the logarithmic regret obtained by den Boer and Zwart (2015). These authors study a parametric model where the unknown demand function is characterized by two parameters. It is shown that 'learning takes care of itself'; a near-myopic policy with full emphasis on 'exploitation' performs very well and learns the parameters 'on the fly'. This property is not true in our case; a myopic policy that does not pay careful attention to exploring all actions sufficiently often would incur a loss that grows linearly with n. The need to put more emphasis on 'exploration' naturally induces a higher regret rate.

An interesting direction for future research is to see whether the $n^{2/3}$ rate of Theorem 1 can be improved, and to prove a lower bound on the (worst-case) regret achieved by any policy.

5 Extension

In this section, seasons are allowed to be non-identical: season length and initial inventory are allowed to vary across different seasons. Two strategies are studied: (i) strategy σ'' merely extends σ to allow non-identical seasons; when seasons are identical, the two strategies coincide; (ii) strategy σ' extends σ in the same sense, but also modifies it by requiring policy updates during exploitation. In our numerical results (all of which use identical seasons), σ' outperforms σ for modest time horizons; this is what motivates it. We prove a performance guarantee analogous to that in Theorem 1 for both σ' and σ''.

We revise the model of Sect. 2 as follows. At the beginning of any season s, the inventory is replenished to $x_s \in \mathbb{N}$, and the seasonal remaining time is $T_s \in \mathbb{N}$. The terms *initial inventory* and *season length*, when speaking of season s, refer to x_s and T_s, respectively. The sequences of season lengths and initial inventories are bounded: $\overline{T} := \sup_{j\in\mathbb{N}} T_j < \infty$, and $\overline{x} := \sup_{j\in\mathbb{N}} x_j < \infty$. All seasons share the same set of feasible prices, $\{p_1, \ldots, p_\kappa\}$, and vector of purchase probabilities, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_\kappa)$. The inventory dynamics are

$$C_u = \left\{ \begin{array}{ll} x_s & \text{if } u = \sum_{k=1}^{s-1} T_k + 1 \\ \max\{C_{u-1} - D_{u-1}, 0\} & \text{if } \sum_{k=1}^{s-1} T_k + 1 < u \le \sum_{k=1}^{s} T_k \end{array} \right\}, \quad u = 1, 2, \ldots. \tag{20}$$

The regret of a strategy σ over the first n seasons is $\mathcal{R}_n := \mathcal{R}_n(\sigma; \boldsymbol{\lambda}) := \sum_{s=1}^{n} \big(V_s - \mathbb{E}_\sigma[\sum_{u\in U_s} p_{A_u} \min\{C_u, D_u\}]\big)$, where $U_s := \{u \in \mathbb{N} : \sum_{k=1}^{s-1} T_k + 1 \le u \le \sum_{k=1}^{s} T_k\}$ is the set of time periods belonging to season s, $\mathbb{E}_\sigma$ denotes expectation under σ, and $V_s := V_{T_s,x_s}$ is the optimal value for season s under full-information, as defined in Sect. 2. The strategy with policy updates is:

STRATEGY $\sigma'(\tau)$.

Step 1 (Initialization). Let $\tau \in \mathbb{N}$, $\tau \le n$.
Step 2 (Initial Exploration). For all $u = 1, \ldots, \sum_{k=1}^{\tau} T_k$, set $A_u := \min\{\arg\min \{N_i(u-1) : 1 \le i \le \kappa\}\}$ if $C_u > 0$, and set $A_u = 0$ if $C_u = 0$.
Step 3a (Estimation). For each $s \in \{\tau+1, \ldots, n\}$ and $i = 1, \ldots, \kappa$, let $N_{s-1,i} := \sum_{j=1}^{s-1} \sum_{u\in U_j} \mathbb{1}_{[A_u=i]}$ be the count of time periods in the first $s-1$ seasons during which the price was p_i, and let $S_{s-1,i} := \sum_{j=1}^{s-1} \sum_{u\in U_j} \mathbb{1}_{[A_u=i]} D_u$ be the count of sales obtained in these periods. Set $\widehat{\lambda}_{s,i} := S_{s-1,i} N_{s-1,i}^{-1} \mathbb{1}_{[N_{s-1,i}>0]}$ for $i = 1, \ldots, \kappa$ and $\widehat{\boldsymbol{\lambda}}_s := (\widehat{\lambda}_{s,1}, \ldots, \widehat{\lambda}_{s,\kappa})$.
Step 3b (Exploitation). For each $s \in \{\tau+1, \ldots, n\}$, apply the policy $\widehat{\pi}_s = \pi(\widehat{\boldsymbol{\lambda}}_s)$ defined in (4), during season s.

Theorem 2 *Set $\tau_n \asymp (n^2 \log n)^{1/3}$. Then, there exists a finite positive constant K_2 such that, for all $n \ge 2$,*

$$\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \mathcal{R}_n(\sigma'(\tau_n); \boldsymbol{\lambda}) \le K_2 (n^2 \log n)^{1/3}. \tag{21}$$

Proof of Theorem 2 The proof follows that of Theorem 1. Let $n \ge 2$.

Step 1. For all $s \in \{\tau_n + 1, \ldots, n\}$, let $\widehat{\boldsymbol{\lambda}}_{n,s}$ be the estimate obtained in step (3a) of $\sigma'(\tau_n)$, and let $\widehat{\pi}_{n,s}$ be the policy applied in step (3b), with value $V_{n,s} := V^{\widehat{\pi}_{n,s}}_{T_s,x_s}(\widehat{\boldsymbol{\lambda}}_{n,s})$ as defined in (2). Then

$$\begin{aligned} \mathcal{R}_n(\sigma'(\tau_n); \boldsymbol{\lambda}) &\overset{(a)}{\le} \sum_{s=1}^{\tau_n} V_s + \mathbb{E}\left[\sum_{s=\tau_n+1}^{n} \big(V_s - V_{n,s}\big)\right] \\ &\overset{(b)}{\le} \tau_n \overline{x}\,\overline{p} + \sum_{s=\tau_n+1}^{n} 4 T_s \overline{p} \cdot \mathbb{E}\left\|\widehat{\boldsymbol{\lambda}}_{n,s} - \boldsymbol{\lambda}\right\| \end{aligned}$$

$$\leq \tau_n \overline{xp} + n \cdot 4\overline{Tp} \cdot \max_{\tau_n < s \leq n} \mathbb{E} \left\| \widehat{\boldsymbol{\lambda}}_{n,s} - \boldsymbol{\lambda} \right\|, \tag{22}$$

where (a) and (b) are simple extensions to their counterparts in (13).

Step 2. We claim that for any $\delta > 0$,

$$\max_{\tau_n < s \leq n} \mathbb{E} \left\| \widehat{\boldsymbol{\lambda}}_{n,s} - \boldsymbol{\lambda} \right\| \leq \delta + 2\kappa \exp(-2\lfloor \underline{f} \tau_n \rfloor \delta^2), \tag{23}$$

where $\underline{f} := \min_{1 \leq s \leq \tau_n} \min\{T_s, x_s\}/\kappa \geq 1/\kappa > 0$. To prove this, we bound the sample sizes associated to $\widehat{\boldsymbol{\lambda}}_{n,s}$ from below:

$$\min_{\tau_n < s \leq n} N_{s-1,i} \geq N_{\tau_n,i} =: N_i \overset{(a)}{\geq} \lfloor \underline{f} \tau_n \rfloor \quad \text{for each } i, \tag{24}$$

where (a) uses the fact that in each season $s \leq \tau_n$, the inventory is positive during at least $\min_{1 \leq s \leq \tau_n} \min\{T_s, x_s\} = \underline{f}\kappa$ periods, combined with the near-parity of sample sizes ($|N_i - N_j| \leq 1$ for $i \neq j$). Now (23) follows from (24) as in the proof of Proposition 1 , with f replaced by $\underline{f}$. The remainder of the proof mimics that of Theorem 1, step 2, with f replaced by $\underline{f}$. □

We now define the strategy σ'' and state a performance guarantee for it in Corollary 1 below; the proof follows easily from that of Theorem 2.

STRATEGY $\sigma''(\tau)$.

Steps 1-2. Identical to those of $\sigma'(\tau)$.

Step 3 For each season $s = \tau + 1, \ldots, n$, apply the policy $\widehat{\pi} = \pi(\widehat{\boldsymbol{\lambda}}_{\tau+1})$ defined in (4), where $\widehat{\boldsymbol{\lambda}}_{\tau+1}$ is defined as in strategy $\sigma'(\tau)$.

Corollary 1 *Let* $\tau_n \asymp (n^2 \log n)^{1/3}$. *Then, for all* $n \geq 2$, *we have* $\sup_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \mathcal{R}_n(\sigma''(\tau_n); \boldsymbol{\lambda}) \leq K_2 (n^2 \log n)^{1/3}$, *with* K_2 *as in the proof of Theorem* 2.

6 Numerical results

Strategies σ, σ' will be compared with six others, which are all recent and different strategies with proven performance guarantees in particular settings: (1) two strategies based on the fluid approximation in Besbes and Zeevi (2012, Algorithm 1, Section 3.1); (2) two strategies that adapt the upper-confidence-bound approach of Babaioff et al. (2015); (3) Thompson sampling with inventory updating (Ferreira et al. 2018, Algorithm 2); and (4) the method of den Boer and Zwart (2015), adapted for a finite price set.

The next section elaborates these alternatives.

6.1 Alternative strategies

Fluid-based strategies σ_F and σ'_F These strategies are inspired by Besbes and Zeevi (2012, Algorithm 1, Section 3.1). With $\boldsymbol{\lambda}$ momentarily assumed known, consider the

linear program

$$\max\Big\{\sum_{i=1}^{\kappa} p_i\lambda_i t_i\,:\,\sum_{i=1}^{\kappa}\lambda_i t_i\le x,\ \sum_{i=1}^{\kappa} t_i\le T,\ t_i\ge 0,\ i=1,\dots,\kappa\Big\},\qquad(25)$$

and define a *fluid-optimal* policy $\pi_F=\pi_F(\boldsymbol{\lambda})$ as follows. Let $\boldsymbol{t}:=(t_1,\dots,t_\kappa)=\boldsymbol{t}(\boldsymbol{\lambda})$ be an extreme-point optimal solution of the linear program. Let m be the number of positive elements of $\boldsymbol{t}$, and note m is either one or two. If $m=1$, then apply, until stock-out or the season's end, the price that corresponds to the unique component of $\boldsymbol{t}$ that is positive. In case that $m=2$, let i_1 and i_2 be the indices of the two positive elements of $\boldsymbol{t}$, ordered in increasing revenue rate: $\lambda_{i_1}p_{i_1}\le\lambda_{i_2}p_{i_2}$, and apply, until stock-out or the season's end, price p_{i_1} for the first $[t_{i_1}]$ periods and price p_{i_2} otherwise. The ordering "first p_{i_1}, then p_{i_2}" is chosen because it performed somewhat better than the reverse one (first p_{i_2}, then p_{i_1}) in our small-inventory cases ($x=10$), while the two were indistinguishable when $x=100$. Besbes and Zeevi (2012) tacitly prove that the ordering is immaterial (in their model) as inventory grows large: it appears neither in Algorithm 1 there, nor in the associated regret bound (Besbes and Zeevi 2012, Theorem 1). We now define two explore-then-exploit strategies for a horizon of n seasons:

STRATEGY $\sigma_F=\sigma_F(\tau)$ AND STRATEGY $\sigma'_F=\sigma'_F(\tau)$

Step 1. During the first τ seasons, price to learn, maintaining near-parity of sample sizes (similar as under σ). Let $\widehat{\boldsymbol{\lambda}}$ be the estimate of $\boldsymbol{\lambda}$ based on the history up to the end of season τ.
Step 2, Strategy σ_F. For each season $s=\tau+1,\dots,n$, apply the counterpart of π_F in which $\boldsymbol{\lambda}$ is replaced by $\widehat{\boldsymbol{\lambda}}$.
Step 2, Strategy σ'_F. For each season $s=\tau+1,\dots,n$:

(a) Let $\widehat{\boldsymbol{\lambda}}_s$ be the estimate of $\boldsymbol{\lambda}$, based on the history up to season s.
(b) Apply the counterpart of π_F in which $\boldsymbol{\lambda}$ is replaced by $\widehat{\boldsymbol{\lambda}}_s$.

Note that σ_F fixes a single policy throughout the exploitation phase, whereas σ'_F re-estimates and updates the policy in each season. In choosing τ, we considered the following variants: $\tau=\lceil c_\tau(n^2\log n)^{1/3}n^{i/10}\rceil$ for $i\in\{-2,-1,0,1,2\}$, with c_τ as in (18). For $n=10^6$ (the largest value we considered), the regret was similar for $i\in\{-2,-1,0\}$, and larger otherwise. We therefore choose $i=0$, i.e., set $\tau=\tau_n$ as in (18), and we claim that the performance and inconsistency reported below are not artefacts of having chosen τ poorly as a function of n.

Fluid-based, upper-confidence-bound strategies σ_U and σ'_U Babaioff et al. (2015) approximate the total revenue over a season as $r(p)=r(p;x,T)=r(p;x,T,\lambda(\cdot)):=p\cdot\min(x,T\lambda(p))$, where the price p lies in a continuous domain, and $\lambda(p)$ is the associated purchase probability. They pursue a fixed-price policy that prices at the maximizer of $r(\cdot)$. Their method is asymptotically optimal in their setting, and uses an upper confidence bound (UCB) for each $r(p)$, for p in an appropriate finite set that is asymptotically dense in the continuous domain. Translating this approach

to the finite-price setting, we seek to price at

$$i^* := \min \arg\max_{i \in \mathcal{A}} r(p_i) \tag{26}$$

through their UCB method detailed below. Babaioff et al. (2015, Section 4) also mention a "tempting", dynamic alternative, in which, *at each time u*, the total *remaining* revenue in the season is approximated by $r_u(p_i) = r(p_i; C_u, t_u) := p_i \cdot \min(C_u, t_u \lambda_i)$ (where C_u is the remaining inventory and t_u is the remaining time), and one aims to price at the maximizer of $r_u(\cdot)$, via the same UCB method. To implement both these variants, we use the upper confidence bounds in Babaioff et al. (2015), as follows: let $N_i(u)$ denote the number of periods before u in which the chosen price was equal to p_i; let $S_i(u)$ denote the total sales obtained during these periods; and define

$$\begin{aligned} \widehat{\lambda}_{u,i} &= \mathbb{1}_{[N_i(u)>0]} \frac{S_i(u)}{N_i(u)} + \mathbb{1}_{[N_i(u)=0]}, \\ \rho_{u,i} &= \frac{\alpha}{N_i(u)+1} + \sqrt{\frac{\alpha \widehat{\lambda}_{u,i}}{N_i(u)+1}} \quad \text{for } \alpha := \log(T), \\ I_{u,i} &= p_i \cdot \min\left\{x, T(\widehat{\lambda}_{u,i} + \rho_{u,i})\right\}. \end{aligned} \tag{27}$$

Here, $I_{u,i}$ is an upper confidence bound on $r(p_i)$, with the radius $\rho_{u,i}$ motivated in Babaioff et al. (2015). In addition, define an index $I'_{u,i}$ as in (27), with x and T replaced by C_u and t_u respectively; this index is an upper confidence bound on $r_u(p_i)$. We now define two strategies for n seasons of length T.

STRATEGY σ_U For all $u = 1, \ldots, nT$, set $A_u = \min \arg\max_{1 \le i \le \kappa} I_{u,i}$ if $C_u > 0$ and set $A_u = 0$ if $C_u = 0$.

STRATEGY σ'_U For all $u = 1, \ldots, nT$, set $A_u = \min \arg\max_{1 \le i \le \kappa} I'_{u,i}$ if $C_u > 0$, and set $A_u = 0$ if $C_u = 0$.

Thus, both these strategies seek to charge a price that maximizes the UCB on corresponding expected revenue; in case of a tie, the smallest maximizing price is selected.

Thompson-sampling strategy σ_T This strategy is an adaptation of Ferreira et al. (2018, Algorithm 2), which is based on Bayesian estimation of $\boldsymbol{\lambda}$, and, according to the authors, 'addresses the challenge of balancing the exploration-exploitation tradeoff under the presence of inventory constraints'. Following them, the prior distribution on $\boldsymbol{\lambda}$ consists of independent Uniform(0,1) marginals; and since $\boldsymbol{\lambda}$ is constant over seasons, it is natural that we apply their (Bayesian) estimator to the data from all past time periods.

STRATEGY σ_T Repeat the following steps for all $u = 1, \ldots, nT$:

> *Sample Demand.* For each $i = 1, \ldots, \kappa$, let $\tilde{\lambda}_i$ be an independent sample from the Beta$(S_i(u)+1, N_i(u) - S_i(u) + 1)$ distribution, where $N_i(u)$ denotes the number of periods before u such that the chosen price was p_i, and $S_i(u)$ denotes the total sales obtained during these periods.

Price. Let $\boldsymbol{t} := (t_1, \ldots, t_\kappa)$ be an optimal solution to the linear program

$$\max \Big\{ \sum_{i=1}^{\kappa} p_i \tilde{\lambda}_i t_i : \sum_{i=1}^{\kappa} \tilde{\lambda}_i t_i \le C_u/t_u,\ \sum_{i=1}^{\kappa} t_i \le 1,\ t_i \ge 0,\ i = 1, \ldots, \kappa \Big\}, \quad (28)$$

where C_u and t_u denote the season's remaining inventory and remaining time in period u, respectively. Set A_u randomly to one of $1, 2, \ldots, \kappa, 0$ with respective probabilities $t_1, t_2, \ldots, t_\kappa, 1 - \sum_{i=1}^{\kappa} t_i$.
Update history. Observe the demand D_u and update the history.

Parametric strategy σ_P This strategy is our adaptation of den Boer and Zwart (2015) to the finite-price setting. Its basis is the assumption that any price p entails the purchase probability $\lambda(p) = \eta(\beta_1 + \beta_2 p)$, where $\eta(z) := \exp(z)/(1 + \exp(z))$, and $\beta := (\beta_1, \beta_2)$ are unknown parameters. By the conditional independence in our model,

$$\begin{aligned}(D_u | A_u = p) \sim &\text{Bernoulli } (\lambda(p)), \text{ independently of} \\ &\{A_1, \ldots, A_{u-1}, D_1, \ldots, D_{u-1}\}, \text{ for all } u = 1, 2, \ldots. \end{aligned} \quad (29)$$

This is a Generalized Linear Model (GLM) with (canonical) link function $\eta(\cdot)$; thus, maximum-likelihood estimates of β are computable by standard methods.

STRATEGY σ_P Repeat the following steps for all $u = 1, \ldots, nT$:

Estimate the purchase probabilities. Compute $\beta_{u-1,j}$. as a maximum-likelihood estimate of β_j $(j = 1, 2)$ under the GLM (29) as of time u, i.e., based on the data $\{A_1, \ldots, A_{u-1}, D_1, \ldots, D_{u-1}\}$. Let $\widehat{\lambda}_{u,a} := \eta(\widehat{\beta}_{u-1,1} + \widehat{\beta}_{u-1,2} p_a)$ for $a \in \{1, \ldots, \kappa\}$ be the estimated probabilities.
Price. Let $\widehat{\pi}_u$ be the optimal action, defined as in (4) with probabilities there being the estimates $\widehat{\lambda}_{u,a}$ computed above. Set $A_u = \widehat{\pi}_u$, except only if $\widehat{\pi}_u$ is such that, during the completed current season, no price-dispersion occurs (i.e., $t_u = 1$, $C_u = 1$, and setting $A_u = \widehat{\pi}_u$ would make the actions, $A_{u'}$ for all u' of the completed season, equal); in this case only, set A_u by altering $\widehat{\pi}_u$ to the nearest action towards the mid-point of the price domain.
Update history. Observe the demand D_u and update the history.

6.2 Consistency

A strategy is said to be *consistent* if its relative regret converges to zero as $n \to \infty$ (equivalently, its regret is $o(n)$) uniformly over $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$. Strategies σ and σ' are consistent, by Theorem 1 and Theorem 2, respectively. In contrast, all six alternative strategies may fail to be consistent.

Inconsistency of σ_F **and** σ'_F Let $V^F = V^F(\boldsymbol{\lambda})$ denote the expected per-season revenue of policy π_F. Loosely speaking, these strategies incur revenue losses of two types relative to pricing optimally (i.e., using π^* in all seasons): (i) the loss $V - V^F = V(\boldsymbol{\lambda}) - V^F(\boldsymbol{\lambda})$; (ii) the loss due to not knowing π_F exactly. More formally, suppose that (i) for some $\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$ we have $V(\boldsymbol{\lambda}_0) - V^F(\boldsymbol{\lambda}_0) > 0$; and (ii) the value

of the policy applied in any exploitation season (the counterpart of π_F) does not exceed $V^F(\boldsymbol{\lambda}_0)$, and $\tau_n/n \to 0$. Then, the exploitation phase incurs a loss at least $(n-\tau_n)(V-V^F)$, which is $\Omega(n)$ (since $V - V^F > 0$). In this setting, we have $\liminf_{n\to\infty} \mathcal{R}_n(\cdot;\boldsymbol{\lambda}_0)/(nV) \geq 1 - V^F(\boldsymbol{\lambda}_0)/V(\boldsymbol{\lambda}_0) > 0$ for these two strategies; thus, the right side is a fundamental lower bound on the relative regret, and we call it the *fluid gap*.

Inconsistency of σ_U and σ'_U Let π_U denote the (single-fixed-price) policy that prices at p_{i^*}, where i^* is defined in (26), and let $V^U = V^U(\boldsymbol{\lambda})$ denote its expected revenue, defined in (2). Assuming that there exists $\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$ such that $V^U(\boldsymbol{\lambda}_0) < V(\boldsymbol{\lambda}_0)$, and that the (expected) per-season revenue under σ_U is at most $V^U(\boldsymbol{\lambda}_0)$ for all seasons that are large enough (a reasonable assumption), then we have lack of consistency: $\liminf_{n\to\infty} \mathcal{R}_n(\sigma_U;\boldsymbol{\lambda}_0)/(nV) \geq 1 - V^U(\boldsymbol{\lambda}_0)/V(\boldsymbol{\lambda}_0) > 0$. Thus, the right side is a fundamental lower bound on the relative regret, and we call it the *fixed-price gap*. Strategy σ'_U does not lend itself to a similar argument (since the functions $r_u()$ involve the stochastic process $\{(C_u, t_u) : u \geq 1\}$, which is difficult to analyze). Since it is indifferent to the MDP, we expect that, for some $\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$, its asymptotic per-season revenue is smaller than $V(\boldsymbol{\lambda}_0)$, i.e., that it is inconsistent; our numerical results below confirm this.

Inconsistency of σ_T The guarantee in Ferreira et al. (2018, Theorem 2) is inconsequential in our setting for two reasons: (i) their upper bound is on Bayesian regret, while ours is on worst-case regret (over all possible values of $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$) and (ii) by the boundedness of season lengths, the Bayesian regret need not vanish as the season index $n \to \infty$. Since σ_T is indifferent to the MDP, we expect that, for some $\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$, its asymptotic per-season revenue is smaller than $V(\boldsymbol{\lambda}_0)$, i.e., that it is inconsistent; our numerical results below confirm this.

Inconsistency of σ_P This strategy runs the risk of *model misspecification* discussed earlier: if the demand function cannot be appropriately approximated by the assumed parametric model, then, even with an abundance of sales data, the action it prescribes may differ from the optimal one (entailing an asymptotic per-season revenue smaller than the optimum). Our numerical results below confirm this, and include cases where the inconsistency gap is large.

6.3 Numerical study

Main part: regret with emphasis on the effect of n We compare the performance of σ, σ' with that of the six alternatives σ_F, σ'_F, σ_U, σ'_U, σ_T, and σ_P. We consider identical seasons (Sect. 2) and the following demand functions:

- *Step function*: $\lambda_1(p) = \mu_i$ whenever $p_{i-1} \leq p < p_i$, where $p_i = i/3$ for $i = 0, 1, 2, 3$; and $\mu_i = \exp(-\theta y_i)$, where $\theta = -\log(1/100) = 4.6052$ and $y_i = (p_i + p_{i-1})/2$ (resulting in $\mu_1 = 0.4642$, $\mu_2 = 1/10$, $\mu_3 = 0.0215$).
- *Linear*: $\lambda_2(p) = 1 - p$.
- *Logistic (Logit)*: $\lambda_3(p) = \eta(\beta_1 - \beta_2 p)$, where $\eta(z) := \exp(z)/(1+\exp(z))$, and such that $\lambda_3(1) = 1 - \lambda_3(0) = 1/100$ ($\beta_1 = 4.5951$, $\beta_2 = 9.1902$).
- *Exponential*: $\lambda_4(p) = \exp(-\theta p)$, where $\theta = -\log(1/100)$ (resulting in $\lambda_4(0) = 1$, $\lambda_4(1) = 1/100$).

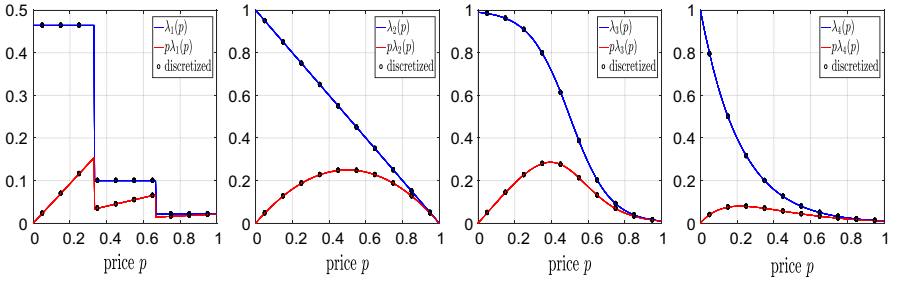

Fig. 1 The demand and revenue vectors, and the generating continuously-supported analogs for the step, linear, logit, and exponential cases

The step function is chosen to have a small number of discontinuities of substantial size; all other three demand functions are continuous. (See den Boer and Keskin (2019) for several practical applications where demand discontinuities may arise). The number of prices is set to $\kappa = 10$, and the price points to $p_i := (i - 0.5)/\kappa$ for $i = 1, 2, \ldots, \kappa$. This results in the demand vectors $\boldsymbol{\lambda}_i = [\lambda_i(p_1), \ldots, \lambda_i(p_\kappa)]$ for $i \in \{1, 2, 3, 4\}$. Figure 1 depicts the demand vectors $\boldsymbol{\lambda}_i$; the revenue vectors $\{p_j \lambda_i(p_j)\}_{j=1}^{10}$; and the underlying continuously-supported analogs ($\lambda_i(p)$ and $p\lambda_i(p)$ for $p \in [0, 1]$), for all i. Onwards, the four demand vectors are treated in a unified manner.

Inventory is set at the levels $x_1 = 10$ and $x_2 = 100$. We examine the effect of demand strength systematically as follows. For each demand vector $\boldsymbol{\lambda}_i$, $i \in \{1, 2, 3, 4\}$, let p_i^U be the revenue-rate maximizing price (i.e. $\lambda_i(p_i^U) p_i^U = \max_j \lambda_i(p_j) p_j$). We want the mean demand, when price p_i^U is applied throughout the season, to be as close as possible to the inventory x_k times a demand-strength factor c_j; this is achieved by setting the season length as

$$T_{k,i,j}(x) = [c_j x_k / \lambda_i(p_i^U)]. \tag{30}$$

We set $c_j = (3/4) \cdot 2^{j-1}$ for $j = 1, 2, 3$ to create scenarios of low, medium, and high demand, respectively.

The planning horizon (number of seasons) n varies along powers of 10 from small (10) to large (10^6).

Part 2 In this part we keep the inventory and demand vectors as previously; and we modify the range of demand strength, making it far wider relative to the main experiment: season lengths are set as usual by (30), but now across $c_j = (3/4) \cdot 2^j$ for $j = -1, 0, 1, \ldots, 6$; thus, for $j = -1$ the mean demand is very weak (37.5% of inventory), while for $j = 6$ the mean demand is extremely strong (48 times the inventory). For each resulting case, we report: (a) the fluid gap and the fixed-price gap (defined in Sect. 6.2); and (b) the estimated relative regret of selected strategies for $n = 10^2$ and $n = 10^4$, with the latter serving as a large-n example (estimation details are discussed in Sect. 6.4 below).

Computing cost We study the strategies' computing cost, defined as the time spent computing the price (A_u for all relevant u), as measured within our matlab code via the recommended method, commands *tic* and *toc*. This is done in an experiment where

variable independent factors are the inventory, taking values $x_1 = 50$, $x_2 = 100$, and $x_3 = 200$; the demand vector $\boldsymbol{\lambda}_i$, $i \in \{1, 2, 3, 4\}$; and the demand strength, taking values $c_j = (3/4) \cdot 2^j$ for $j \in \{0, 1, 2, 3\}$; setting season lengths as in (30), the mean demand is 75%, 150%, 300%, and 600% of the inventory, for each k and i. For the dependence of the cost on n, our numerical experience suggests a clear distinction between the 'update' strategies (σ', σ'_F, σ_U, σ'_U, σ_T and σ_P) and the others (σ and σ_F); for the former, the cost is (nearly) linear in n, that is, very close to Cn, where C is the expected cost per season and is strategy-specific, while for the latter it does not change substantially with n. Now, the issues of main interest are (i) the dependence of the update-type strategies' C on the inventory x and season length T; and (ii) the cross-strategy comparison of these C's. We answer these questions based on $n = 100$ seasons; based on unreported results with $n = 10^3$, we expect that these answers would be essentially unchanged for all $n \geq 100$.

6.4 Results

We sometimes refer to the eight strategies generically as σ_ℓ, $\ell = 1, \ldots, 8$.

Main results: relative regret and consistency For each strategy we have 24 cases generated by inventory $x \in \{10, 100\}$; demand vector $\boldsymbol{\lambda}_i$ for $i = 1, 2, 3, 4$; demand level c_j, $j = 1, 2, 3$ (low, medium, and high). For each ℓ and each $n = 10^k$ with $k = 1, \ldots, 6$ (except for σ_P with $k = 5, 6$) we compute an estimate of $\mathcal{R}'_n = \mathcal{R}'_n(\sigma_\ell)$ that is as accurate as possible subject to our computer-time constraints (see details in paragraph *'Estimation and accuracy'* below). These estimates are denoted $\widehat{\mathcal{R}'_n} = \widehat{\mathcal{R}'_n}(\sigma_\ell)$ and are reported in Figs. 2 and 3, for $x = 10$ and $x = 100$ respectively.

Estimation and accuracy For each ℓ and each case, a case-specific number $n_{\text{rep}} = n_{\text{rep}}(\sigma_\ell, x, i, j, n)$ of independent simulations (replications) (of n seasons) is run; each replication yields one sample value of the revenue loss relative to the optimum, $nV_{x,T}$; this, divided by this optimum, is a sample value of $\mathcal{R}'_n = \mathcal{R}'_n(\sigma_\ell)$. The estimate $\widehat{\mathcal{R}'_n}$ is the average of these n_{rep} samples; its relative error (inverse accuracy) is $\text{SE}(\widehat{\mathcal{R}'_n})/\widehat{\mathcal{R}'_n}$, where $\text{SE}(\widehat{\mathcal{R}'_n})$ is the sample standard deviation divided by $\sqrt{n_{\text{rep}}}$. We do not simulate σ_P for large n ($n \in \{10^5, 10^6\}$), because the simulation cost is exceptionally high, except for the two cases where $x = 10$, demand vector is $\boldsymbol{\lambda}_1$, and demand strength is medium or high (which demonstrate the large inconsistency gap); points missing in the figures are due to this choice. Except for σ_T and σ_P, the accuracy is good for all n (relative error less than 5%; for σ and σ_F, less than 2%). The accuracy decreases somewhat for σ_T and σ_P as n increases, but only when $\widehat{\mathcal{R}'_n}(\cdot)$ is very small, in which case our comparisons are unaffected, even if we replace this estimate by the normal-based, 95%-confidence lower or upper bound. This limitation is unavoidable and due to the excessive simulation cost (for example, for $n = 10^6$, a single replication of σ_T for $x = 100$, $i = 1$, and $j = 1$ requires 3.5×10^5 seconds; and σ_P would require much more).

Part-2 results Figures 5 and 6 show, for $x = 10$ and $x = 100$ respectively, the two gaps and the relative regrets for $n = 10^2$ and $n = 10^4$. Since these gaps explain the large-n regret of π_F and π_U, we examine them more carefully. We define the *fluid (LP) slack* and the *fixed-price (FP) slack* as the fraction of inventory that is unused under

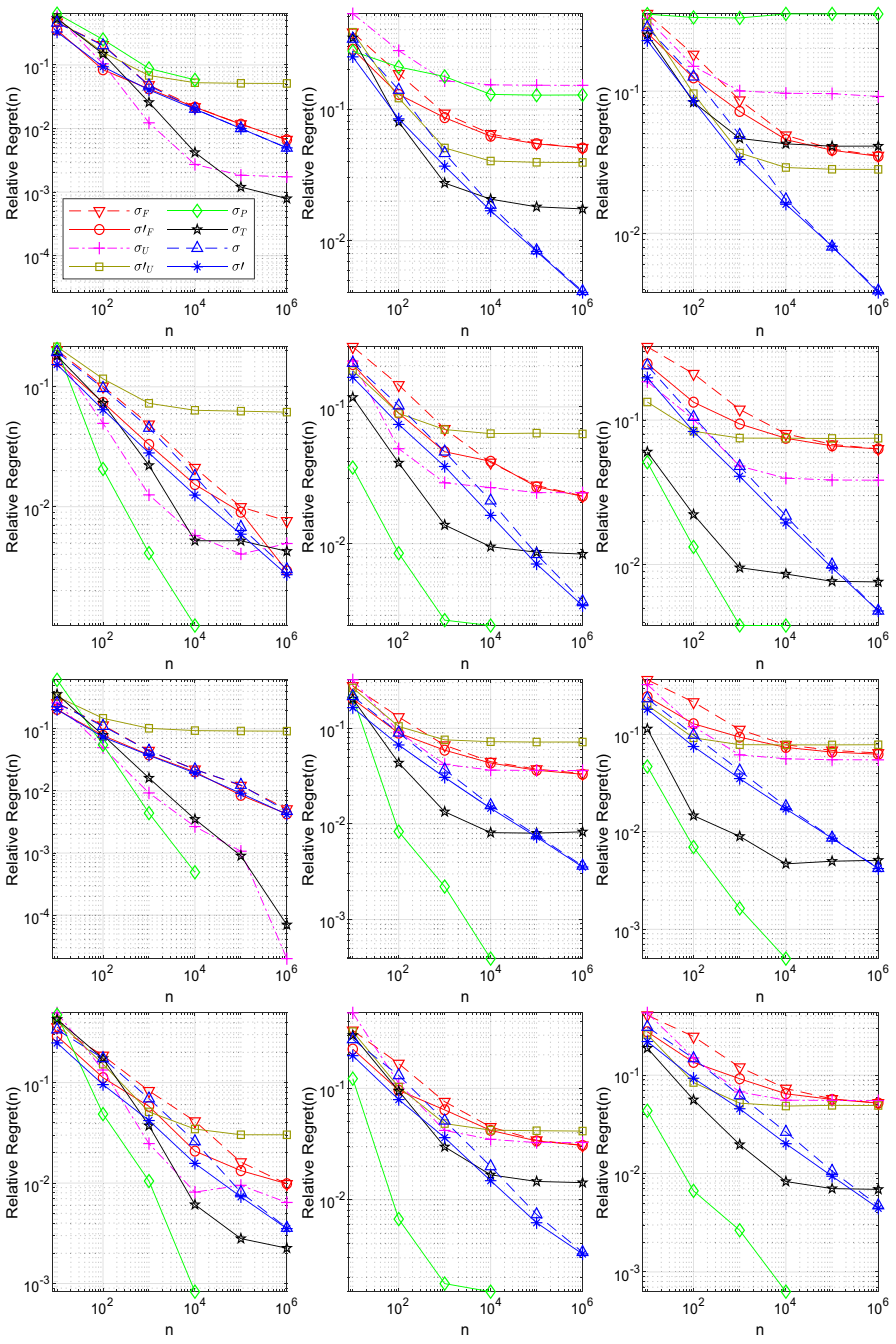

Fig. 2 The strategies' estimated relative regrets, $\widehat{\mathcal{R}'_n}(\cdot)$, (vertical axis) as functions of the number of seasons n (horizontal axis) for $x = 10$. The figure in row i, column j corresponds to demand vector $\boldsymbol{\lambda}_i$ and season length $T_{1,i,j}$ (low, medium, and high demand in the left, center, and right column, respectively)

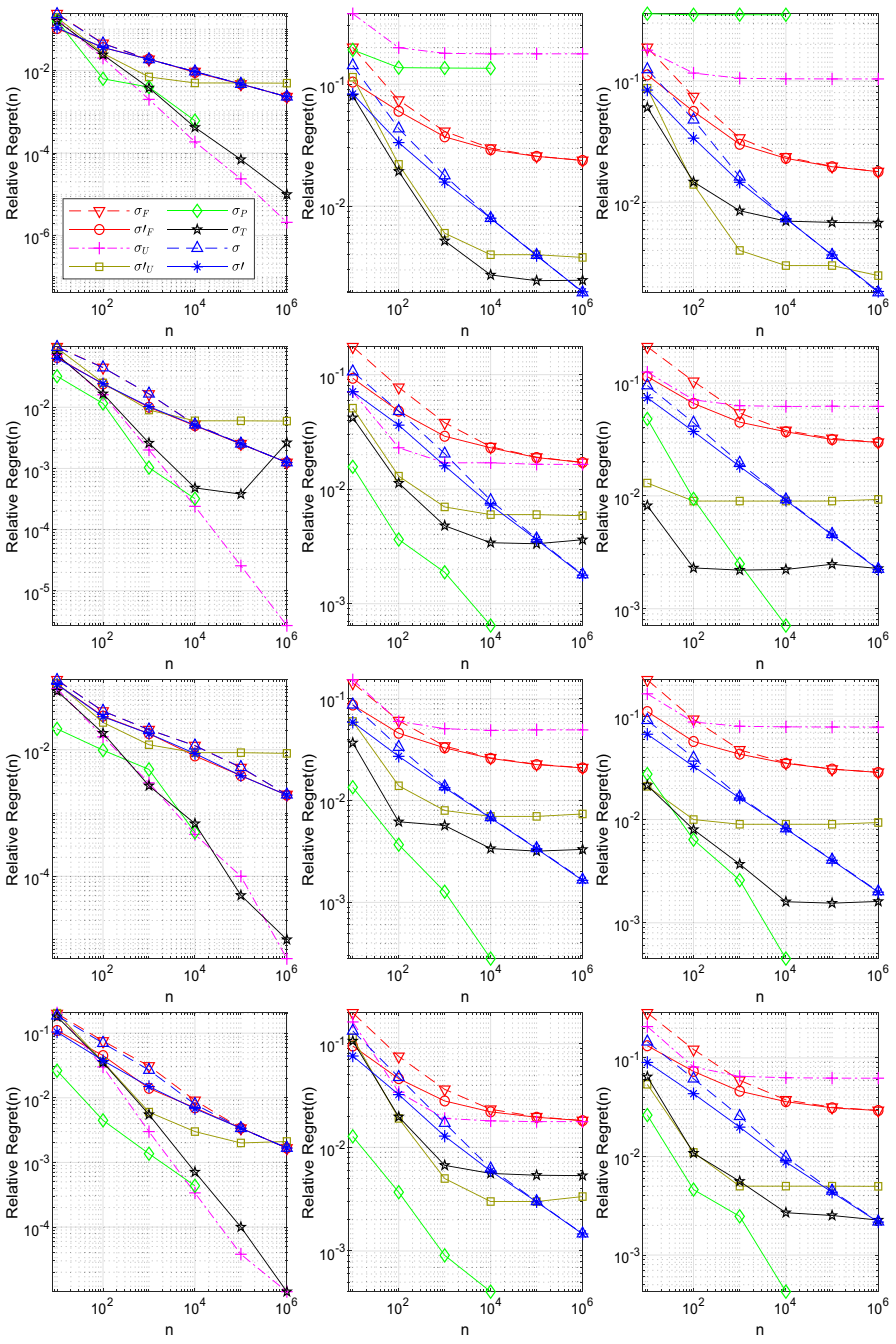

Fig. 3 The strategies' estimated relative regrets, $\widehat{\mathcal{R}'_n}(\cdot)$, (vertical axis) as functions of the number of seasons n (horizontal axis) for $x = 100$. The figure in row i, column j corresponds to demand vector $\boldsymbol{\lambda}_i$ and season length $T_{2,i,j}$ (low, medium, and high demand in the left, center, and right column, respectively)

the optimal solution in the fluid model underlying π_F and π_U, respectively. Figure 7 gives a detailed view of these optimal solutions for each of the 8 cases (two x, four $\boldsymbol{\lambda}$), showing, in each case, the optimal price (or prices in the former case), the slack, and the gap.

Insights obtained are: (a) these gaps are, respectively, tight lower bounds, on the large-n relative regret of σ_F and σ_U; and (b) at the extremes stated above, σ is not competitive, which is explained by the fact that the optimal policy is nearly a fixed-price one, and consequently the gaps nearly vanish. Since these results are secondary, they are presented (and discussed) in the Appendix.

Results on computing cost We have defined $48 = 3 \times 4 \times 4$ design points. With $C^{\ell}_{k,i,j}$ denoting the cost of strategy σ_ℓ at the point (k, i, j) (i.e., with capacity x_k, demand function i, and demand strength c_j), we report the summary statistics over the design, $C^{\ell}_{\min} = \min_{k,i,j} C_{k,i,j}$, $C^{\ell}_{\text{avg}} = \frac{1}{48}\sum_{k=1}^{3}\sum_{i=1}^{4}\sum_{j=1}^{4} C_{k,i,j}$, and $C^{\ell}_{\max} = \max_{k,i,j} C_{k,i,j}$. Moreover, we develop a cost model $C^{\ell}(x, T) \approx 10^{\beta^{\ell}_0} x^{\beta^{\ell}_x} T^{\beta^{\ell}_T}$, and estimate it by allowing random error over the design points:

$$\log_{10} C^{\ell}_{k,i,j} = \beta^{\ell}_0 + \beta^{\ell}_{x_k} \log_{10} x_k + \beta^{\ell}_{T_{k,i,j}} \log_{10} T_{k,i,j} + \epsilon^{\ell}_{k,i,j} \tag{31}$$

where $\epsilon^{\ell}_{k,i,j}$ are random errors. We compute (least-squares) estimates $\hat{\beta}^{\ell}_0$, $\hat{\beta}^{\ell}_x$, and $\hat{\beta}^{\ell}_T$; these imply the model $C^{\ell}(x, T) \approx 10^{\hat{\beta}^{\ell}} x^{\hat{\beta}^{\ell}_x} T^{\hat{\beta}^{\ell}_T}$. For each $\ell = 1, \ldots, 6$, the data $C^{\ell}_{k,i,j}$ and the (estimated) model are visualized in Fig. 4; the model and the summary statistics are reported in Table 1.

6.5 Discussion

This Sect. discusses results and insights from the numerical study. Our main results are Figs. 2 and 3, showing for each $\ell = 1, \ldots, 8$ the (estimated) relative regret $\widehat{\mathcal{R}'_n}(\sigma_\ell)$ as a function of n, with both axes shown in logarithmic scale.

Regret consistency and rate of convergence For σ and σ', the relative regret is nearly a straight line with slope $-1/3$ in all 24 cases, in line with Theorems 1 and 2. For each alternative except σ_P, the relative regret flattens (the slope approaches zero) as n increases in all cases except the low-demand ones, i.e., in all sub-figures except those in the left column. For σ_P, a flattening of the relative regret is evident only for the demand vector $\boldsymbol{\lambda}_1$. Whenever a flattening is evident, the flattened value, i.e., $\widehat{\mathcal{R}'_n}(\sigma_\ell)$ for the largest available n, is a reasonable proxy for the inconsistency (gap), defined as $\lim_{n\to\infty} \mathcal{R}'_n(\sigma_\ell)$ (and also discussed in Sect. 6.2). The size of the gap varies with the case and strategy, and is also discussed below. Crucially, σ and σ' enjoy a consistency and convergence rate that hold uniformly across all cases (inventory, demand vector, demand strength), while no other strategy achieves this uniform consistency, even if some perform very well in some cases.

Inconsistent strategies: the size of the gap and consequences thereof Given some inconsistent strategy σ_ℓ, i.e., with gap $\lim_{n\to\infty} \mathcal{R}'_n(\sigma_\ell) > 0$, each of σ and σ' outperforms σ_ℓ for all $n \geq n_0$, for some $n_0 = n_0(\ell)$. Such $n_0 \leq 10^6$ are often exhibited in these figures; for example, for $x = 10$, $\boldsymbol{\lambda}_1$, and high demand, choosing σ' against σ_T,

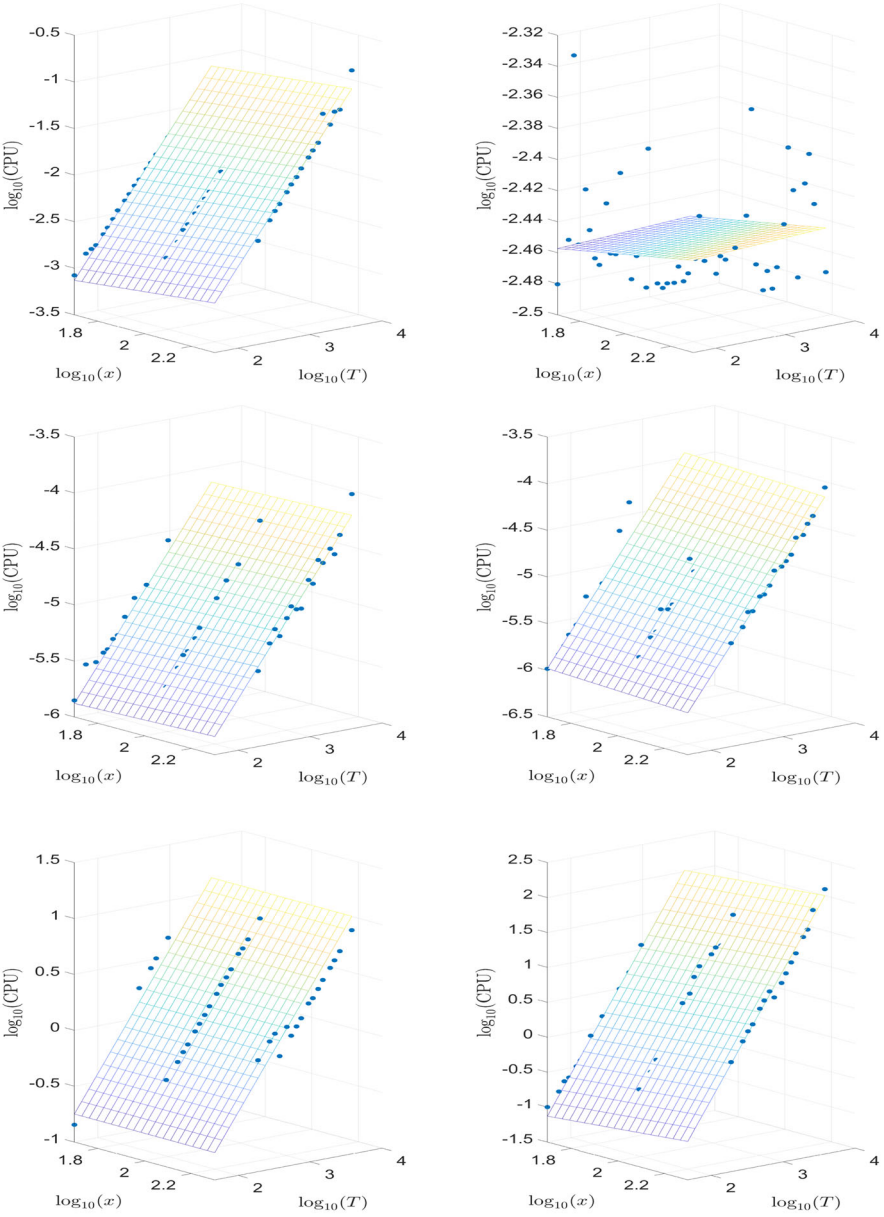

Fig. 4 In each figure, the height is $\log_{10}(C)$, where C is the (per-season) cost; the two axes in the horizontal plane are $\log_{10}(x)$ and $\log_{10}(T)$. Each figure $\ell = 1, \ldots, 6$ shows the cost points $\log_{10}(C^{\ell}_{k,i,j})$ for all (or most) k, i, j. The plane shown is the estimated model. Figures correspond to strategies as follows: σ', σ'_F (first row, from left to right); σ_U, σ'_U (second row); σ_T and σ_P (third row)

Table 1 Summary of the computing cost of key strategies. Model parameters β_0, β_x, and β_T; coefficient of multiple determination R^2; and summary statistics

Strategy	β_0	β_x	β_T	R^2	$C_{\min}$	C_{avg}	$C_{\max}$
σ'	−5.365	0.314	1.040	0.996	7.76e−04	1.93e−02	1.73e−01
σ'_F	−2.508	0.028	0.002	0.050	3.26e−03	3.58e−03	4.62e−03
σ_U	−7.668	0.073	0.955	0.947	1.25e−06	1.72e−05	1.82e−04
σ'_U	−7.584	0.024	0.990	0.971	1.37e−06	1.99e−05	1.76e−04
σ_T	−2.378	−0.012	0.987	0.914	0.143	2.46	10.0
σ_P	−4.393	0.297	1.650	0.988	0.100	14.56	194.0

we see that n_0 is about 100; against any other strategy, n_0 is at most 10. In a number of cases, the gap is apparently small, so we do not obtain $n_0 \leq 10^6$. We see two prominent groups with apparently small gap: (1) all strategies, the low-demand cases (sub-figures in the left column of Figs. 2 and 3); and (2) strategy σ_P, under vectors $\boldsymbol{\lambda}_2$ to $\boldsymbol{\lambda}_4$. These groups are discussed further, respectively, in the two paragraphs that follow. Strategies σ_T and σ_P appear to be strong contenders: for vectors other than $\boldsymbol{\lambda}_1$, and with medium or high demand, their gap is small and their relative regret is smaller than that of σ and σ' for many n (n smaller than the n_0 above). That said, taking σ_T as the contender, we only fail to demonstrate $n_0 \leq 10^6$ in one out of the 16 non-low-demand cases (the case $x = 100$, $\boldsymbol{\lambda}_3$, and high demand).

Effect of extreme-demand scenarios The low-demand cases stand out in that a flattening of the relative regret (of alternatives) is virtually absent; thus, the gaps appear to be notably smaller compared to the other demand levels. Such an effect might occur in both extremes, i.e, extreme-low and extreme-high demand (evidence that includes both extremes is provided by Figs. 5 and 6 in the Appendix). To explain this effect, note that in these extremes the solution to the MDP is nearly a fixed-price policy: under extreme-low demand, it is the revenue-rate maximizer; under extreme-high demand, it is the maximal price $\overline{p}$. In these extremes, since a fixed price is nearly optimal, a focus on the MDP may be unwarranted, and our approach may under-perform.

Strategy σ_P Figures 2 and 3 suggest that the gaps of σ_P under $\boldsymbol{\lambda}_2$ to $\boldsymbol{\lambda}_4$ tend to be small. However, these gaps are hard to measure accurately, not only because they are apparently small, but also because it is impractical to increase n further (as discussed in Sect. 6.4, paragraph *Estimation and accuracy*). Remarkably, σ_P is the worst performer (of all strategies) under demand vector $\boldsymbol{\lambda}_1$ and the best performer otherwise (demand vectors $\boldsymbol{\lambda}_2$ to $\boldsymbol{\lambda}_4$). This contrast is explained by the parametric nature on σ_P: the parametric demand model on which it is based fails to contain a good approximation to $\boldsymbol{\lambda}_1$, while it apparently succeeds in the other cases.

The following two paragraphs discuss the effect of inventory; this discussion is empirical and its importance secondary.

Effect of inventory on small-n regrets Inventory usually has a negative effect on the finite-n relative regrets. Indicatively, we consider σ' and σ_T for $n = 100$. For $x = 10$, $\widehat{\mathcal{R}}'_n(\sigma')$ ranges across the 12 cases from (about) 6.4% to 9.5%; for $x = 100$

the range is 3.3% to 4.4%. For σ_T, for $x = 10$ the range is 1.5% to 17.4%, and for $x = 100$ the range is 0.23% to 2.4%.

Effect of inventory on worst-case gaps of σ_T and σ_P We consider the effect of inventory on the worst (largest across the 12 cases) gaps of σ_T and σ_P, as approximated by $\widehat{\mathcal{R}'_n}(\sigma_\ell)$ for the largest n in each case. For σ_T we obtain: $\max_{i,j} \widehat{\mathcal{R}'_{10^6}}(\sigma_T; \boldsymbol{\lambda}_i, x, j)$ occurs, for both x, at $i = 1$ and $j = 3$ (step- and high-demand); it is about 4.1% for $x = 10$ but only 0.67% for $x = 100$. For σ_P, the maxima also occur in the case $i = 1$, $j = 3$, but the inventory has almost no effect ($\max_{i,j} \widehat{\mathcal{R}'_n}(\sigma_P; x, i, j)$ is 34.9% for $x = 10$ and 35.1% for $x = 100$).

Computing cost Figure 4 and Table 1 show that the cost of each strategy is explained well, over the design range, by the three-parameter model (31), except only for σ'_F, whose main cost is that of solving a linear program (LP), which is insensitive in x and T. Strategy σ' costs well above σ_U and σ'_U, but well below σ_T and σ_P. The cost of σ_P grows faster in T than that of any other strategy, as seen by its larger β_T. To summarize the elements of these costs, we define an *active (time) period* as one such that neither time nor inventory has run out (in the current season). Then, σ' solves one MDP in each season; σ_P solves one MDP, and in addition estimates a Generalized Linear Model, in each active period; σ'_F solves one LP in each season; σ_T solves one LP in each active period; σ_U and σ'_U only require a few elementary numerical operations in each active period.

Summary and insights Our main conclusion is the uniform consistency and convergence rate of our approach (σ and σ') across all possible demand (purchase probability) vectors, a feature not enjoyed by any other strategy; this is evidenced by considering the totality of cases in each of Figs. 2 and 3. The consistency implies that our approach outperforms any inconsistent strategy for all planning horizons $n \geq n_0$, for a suitable n_0. These n_0, often seen in these figures, depend on the strategy's gap, i.e., the limit of its n-season relative regret as $n \to \infty$; the smaller the gap, the larger n_0 tends to be. It is noteworthy that relatively smaller gaps occur in specific cases: (a) for all inconsistent strategies, under extreme-low or extreme-high demand, perhaps because the optimal policy is then almost a fixed-price one; and (b) for strategy σ_P, but only when its parametric model contains a good approximation to the demand vector. Regarding the strategies' computing cost, our approach sits mid-range among the alternatives we considered.

Appendix

We provide the results of 'Part 2' of our experiments and discussion thereof.

Figures 5 and 6 help quantify the relative regret of σ_F and σ_U via the respective gaps (as predicted in Sect. 6.1). In the large-n case, $n = 10^4$ (right column), we see that for all T the (estimated) relative regret of σ_U is nearly indistinguishable from the fixed-price gap $1 - V^U/V$; and that of σ_F is lower-bounded (fairly tightly) by the fluid gap, $1 - V^F/V$.

Since these gaps are essential to the (large-n) regret of these strategies, we examine them more carefully. In Fig. 7, the season length T (and thus the strength of demand) varies over a dense set of points, and reported as functions of T are the following for each of π_U and π_F: the optimal prices, i^* for the former, and i_1, i_2 for the latter; the slack; and the gap.

Effect of T on the gaps For π_U, we see critical T points at which i^* changes upward, entailing a downward jump in demand rate and inventory consumption, and a positive slack; the local maxima of the fixed-price gap are explained by corresponding local maxima of the slack. For π_F, there exist critical T points at which the optimal solution changes; however, the presence of two prices in the solution means that the slack is usually zero.

Effects of inventory x and demand vector $\boldsymbol{\lambda}$ on the worst-in-T gaps For each of the 8 cases (sub-figures) in Fig. 7, we define the worst-in-T gaps $\rho(\pi_F; \boldsymbol{\lambda}, x) := \max_T(1 - V^F(\boldsymbol{\lambda}, x, T)/V(\boldsymbol{\lambda}, x, T))$ and $\rho(\pi_U; \boldsymbol{\lambda}, x) := \max_T(1 - V^U(\boldsymbol{\lambda}, x, T)/V(\boldsymbol{\lambda}, x, T))$. For each $\boldsymbol{\lambda}$ (row), $\rho(\pi_F; \boldsymbol{\lambda}, x)$ decreases with x. For each x (column), the worst (largest) $\rho(\pi_U; \boldsymbol{\lambda}, x)$ (and the largest slacks) occur at $\boldsymbol{\lambda}_1$; the larger sparsity of this vector, where sparsity of $\boldsymbol{\lambda}$ is defined as $\max_{1<i\le\kappa} |\lambda_i - \lambda_{i-1}|$, may be the reason behind the larger slacks and gaps.

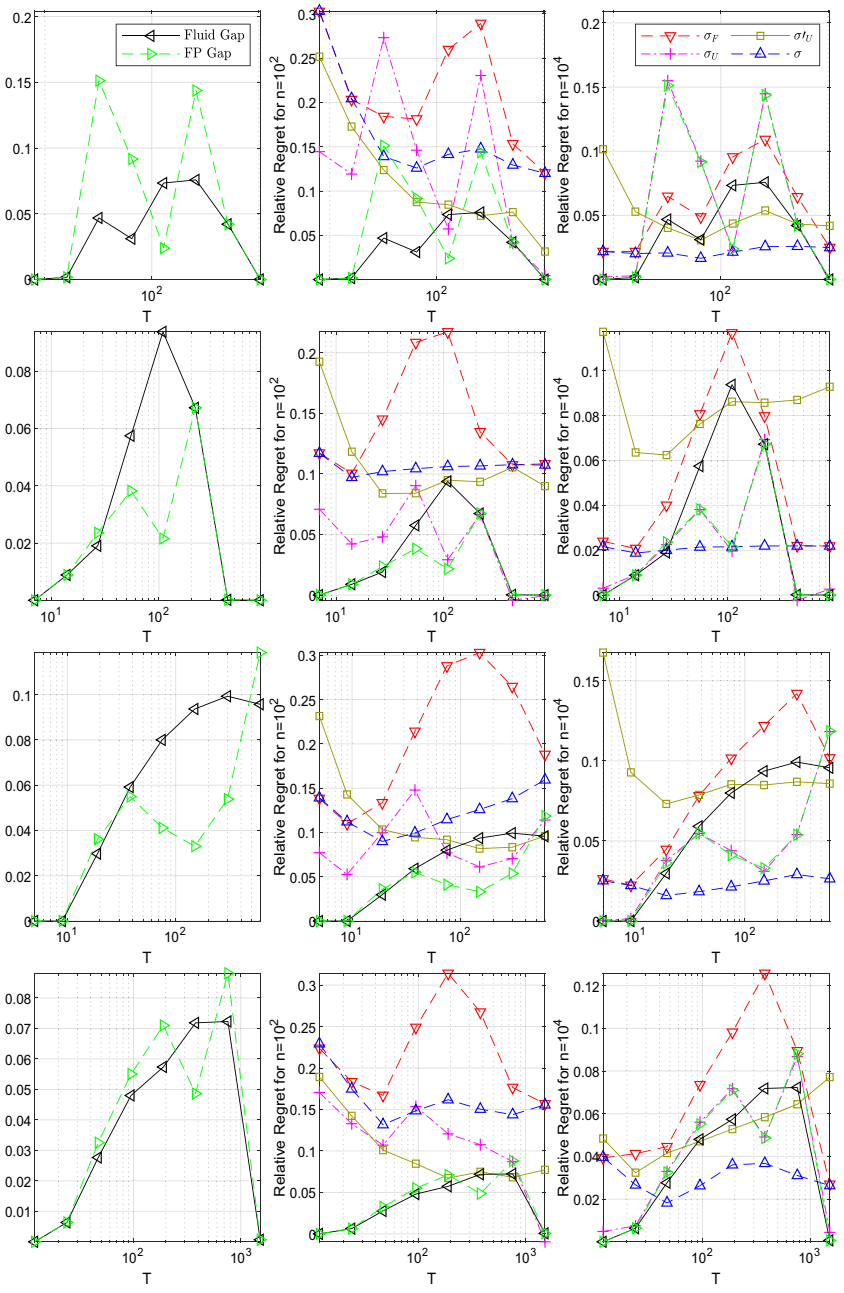

Fig. 5 Gaps and estimated n-season relative regrets of strategies (vertical axis) as functions of season length T (horizontal axis, logarithmic scale) for $x = 10$. The ith row of figures corresponds to demand vector $\boldsymbol{\lambda}_i$. Each figure includes the fluid gap and the fixed-price gap for comparisons. Figures in the center and right column correspond to $n = 10^2$ and $n = 10^4$, respectively

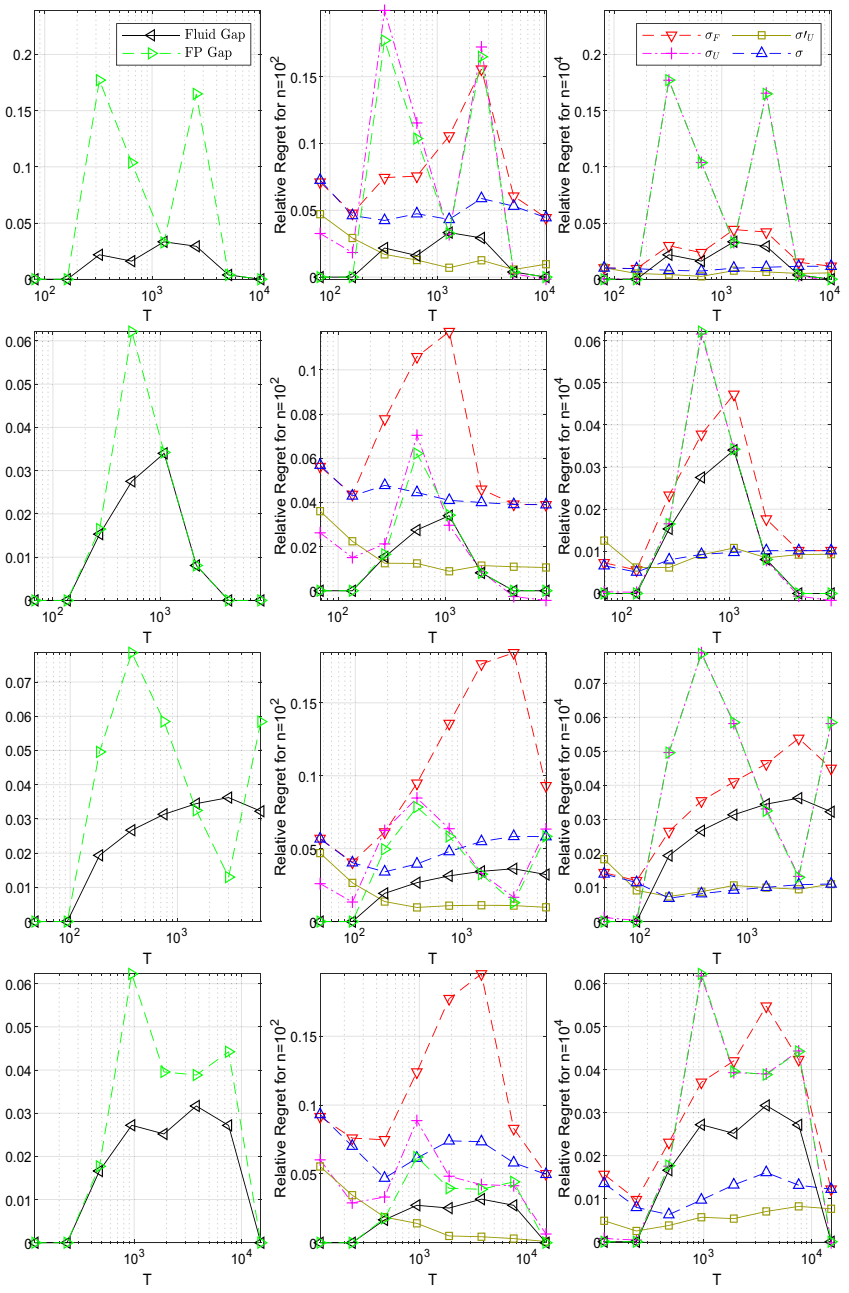

Fig. 6 Gaps and estimated n-season relative regrets of strategies (vertical axis) as functions of season length T (horizontal axis, logarithmic scale) for $x = 100$. The ith row of figures corresponds to demand vector $\boldsymbol{\lambda}_i$. Each figure includes the fluid gap and the fixed-price gap for comparisons. Figures in the center and right column correspond to $n = 10^2$ and $n = 10^4$, respectively

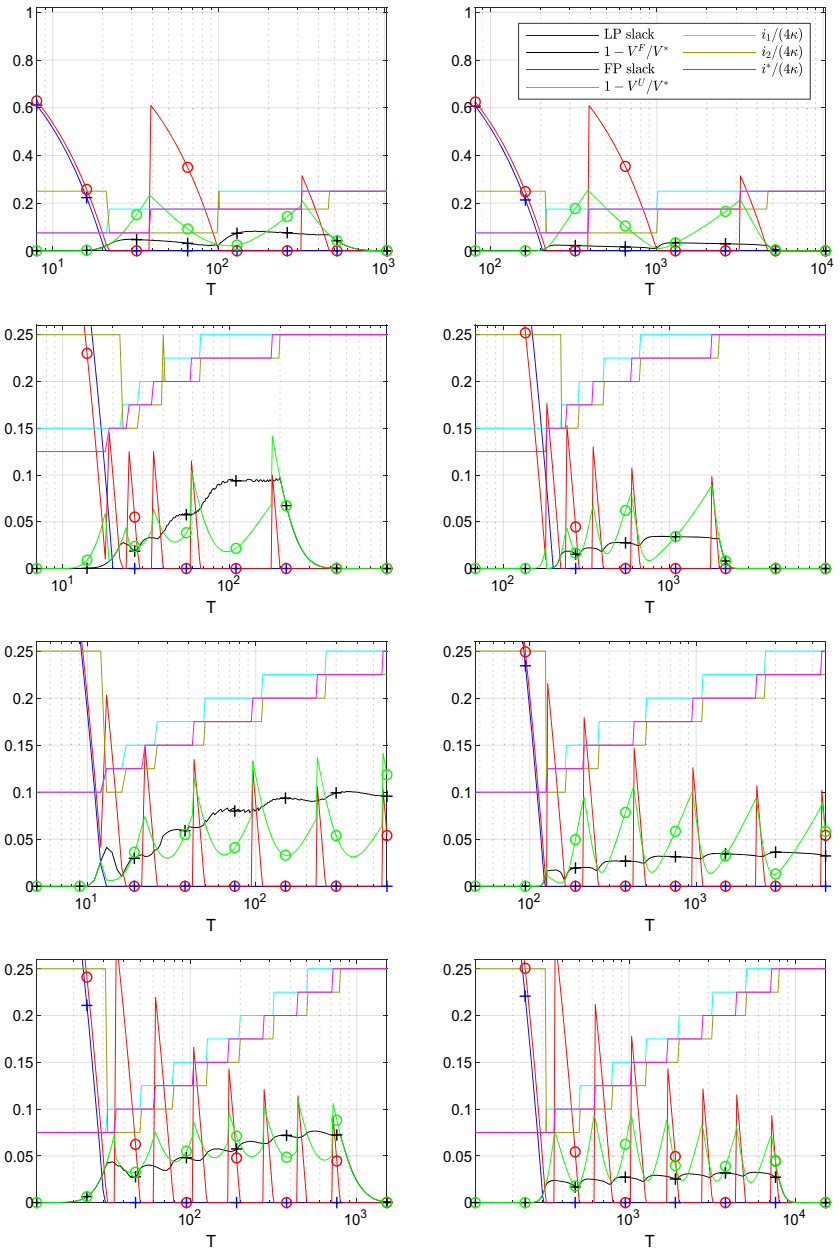

Fig. 7 Properties of policies π_F and π_U as functions of season length T (horizontal axis, logarithmic scale): optimal price indices (divided by $(4\kappa)^{-1}$ for convenient scaling); the slack; and the gap. The ith row of figures corresponds to demand vector $\boldsymbol{\lambda}_i$. The left and right columns correspond to inventory $x = 10$ and $x = 100$ respectively. Symbols '+' and 'o' mark the slack and gap of π_F and π_U, respectively, for the T values in Figs. 5 and 6

References

Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. Oper Res 57(5):1169–1188

Audibert J-Y, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. COLT

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. Mach Learn 47:235–256

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002) The nonstochastic multiarmed bandit problem. SIAM J Comput 32(1):48–77. https://doi.org/10.1137/s0097539701398375

Auer P, Ortner R (2007) Logarithmic online regret bounds for undiscounted reinforcement learning. In: Advances in neural information processing systems 19'. The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0011

Aviv Y, Pazgal A (2005) Pricing of short life-cycle products through active learning. http://www.olin.wustl.edu/faculty/aviv/papers/Pricing_MS_Dec_2005.pdf

Babaioff M, Dughmi S, Kleinberg R, Slivkins A (2015) Dynamic pricing with limited supply. ACM Trans Econ Comput (TEAC) 3(1):1–26

Badanidiyuru A, Kleinberg R, Slivkins A (2013) Bandits with knapsacks. In: 2013 IEEE 54th annual symposium on foundations of computer science. IEEE, pp 207–216

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: risk bounds and near-optimal algorithms. Oper Res 57(6):1407–1420

Besbes O, Zeevi A (2012) Blind network revenue management. Oper Res 60(6):1537–1550

Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. Manage Sci 61(4):723–739

Bitran G, Caldentey R (2003) An overview of pricing models for revenue management. Manuf Serv Oper Manag 5(3):203–229

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. Oper Res 60(4):965–980

Burnetas AN, Katehakis MN (1997) Optimal adaptive policies for Markov decision processes. Math Oper Res 22(1):222–255

Chen B, Chao X, Ahn H-S (2019) Coordinating pricing and inventory replenishment with nonparametric demand learning. Oper Res 67(4):1035–1052

den Boer AV (2015) Dynamic pricing and learning: historical origins, current research, and new directions. Surv Oper Res Manag Sci 20:1–18

den Boer AV, Keskin NB (2019) Discontinuous demand functions: estimation and pricing. Manag Sci 66:4516–4534

den Boer AV, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. Manag Sci 60(3):770–783

den Boer AV, Zwart B (2015) Dynamic pricing and learning with finite inventories. Oper Res 63(4):965–978

Elmaghraby W, Keskinocak P (2003) Dynamic pricing in the presence of inventory considerations: research overview, current practices, and future directions. Manag Sci 49(10):1287–1309

Even-Dar E, Mannor S, Mansour Y (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. J Mach Learn Res 7:1079–2205

Farias VF, Van Roy B (2010) Dynamic pricing with a prior on market response. Oper Res 58(1):16–29

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using thompson sampling. Oper Res 66(6):1586–1602

Gallego G, Topaloglu H (2019) Revenue management and pricing analytics. Springer, New York. https://doi.org/10.1007/978-1-4939-9606-3

Gallego G, van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Manag Sci 40(8):999–1020

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: asymptotically optimal semi-myopic policies. Oper Res 62(5):1142–1167

Lattimore T, Szepesvári C (2019) Bandit algorithms. https://tor-lattimore.com/downloads/book/book.pdf

Lei Y, Jasin S, Sinha A (2014) Near-optimal bisection search for nonparametric pricing with inventory constraint. https://deepblue.lib.umich.edu/handle/2027.42/108717

Lin KY (2006) Dynamic pricing with real-time demand learning. Eur J Oper Res 174(1):522–538

Maglaras C (2011) Dynamic pricing strategies for multiproduct revenue management problems. Wiley encyclopedia of operations research and management science

Talluri K, van Ryzin G (2005) Theory and practice of revenue management. Springer, Berlin
Vogel W (1960) An asymptotic minimax theorem for the two armed bandit problem. Ann Math Stat 31(2):444–451. https://doi.org/10.1214/aoms/1177705907
Wang Z, Deng S, Ye Y (2014) Close the gaps: a learning-while-doing algorithm for a class of single-product revenue management problems. Oper Res 62(2):318–331