Omega 47 (2014) 116–126

Contents lists available at ScienceDirect

Omega

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

Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning

Rupal Rana [a], Fernando S. Oliveira [b,*]

[a] *School of Business and Economic, Loughborough University, UK*
[b] *ESSEC Business School, ESSEC University, Singapore*

ARTICLE INFO

Article history:
Received 26 April 2012
Accepted 15 October 2013
Processed by B. Lev
Available online 13 November 2013

Keywords:
Revenue management
Dynamic pricing
Reinforcement learning
Simulation

ABSTRACT

This paper examines the problem of establishing a pricing policy that maximizes the revenue for selling a given inventory by a fixed deadline. This problem is faced by a variety of industries, including airlines, hotels and fashion. Reinforcement learning algorithms are used to analyze how firms can both learn and optimize their pricing strategies while interacting with their customers. We show that by using reinforcement learning we can model the problem with inter-dependent demands. This type of model can be useful in producing a more accurate pricing scheme of services or products when important events affect consumer preferences. This paper proposes a methodology to optimize revenue in a model-free environment in which demand is learned and pricing decisions are updated in real-time. We compare the performance of the learning algorithms using Monte-Carlo simulation.

1. Introduction

Dynamic pricing is a business strategy that adjusts the product price in a timely fashion in order to allocate the right service, to the right customer, at the right time [1]. It is usually applied when there are uncertainties and seasonality of demand and supply, in an attempt to increase revenue. In particular, with the use of dynamic pricing over the internet the seller is informed about the level of demand in real-time and can price items using the optimal policies computed using historical data. In many industries managers face the problem of establishing a pricing policy that maximizes the revenue from selling a given inventory of items by a fixed deadline. The common characteristics of these industries are that the full inventory of products is available for sale at the beginning of the selling period, no re-ordering is allowed, and the unsold products that remain by the deadline have a zero constant salvage value [2]. This paper is concerned with using reinforcement learning to solve the tactical problem of dynamically pricing these products to maximize the total expected revenue.

Examples of industries that must use dynamic pricing strategies are manufactured goods and services [3,4]. The first category includes goods with limited shelf-life, such as food items, electronic goods or fashion garments, which are usually sold in a finite time-frame, meaning that there is a deadline by which they must be removed from the store. The second category includes the service industries where the service will not generate revenue once the time window for availability has passed, such as airlines, hotels, conference or party facilities, cruise ship holidays, and tickets for trains, theatres, concerts, cinemas and stadiums.

The single product pricing problem addressed in this paper, and originally studied by Gallego and van Ryzin [5], is important as it represents the issues faced in several industries and it is a good test framework for methodological contributions, see [6–15]. The most essential consideration when developing such a pricing policy is demand forecast accuracy. There are two major sources of randomness in demand: the customer arrival rate and customer reservation price. Most academic studies in revenue management assume that the functional relationship between arrival rate and price is known to the decision-maker. This assumption makes the problem more manageable and offers qualitative insights but it is unlikely to provide an optimal policy. In practice it is very rare that the decision-maker has full knowledge of the demand function. For this reason, to impose a structural form on the demand function can lead to model misspecification, resulting in revenue loss.

The main objective of this paper is to propose a model-free approach whereby the transition probabilities between states (i.e., the demand behaviour) are not characterized by a particular distribution. Reinforcement learning techniques, such as *Q*-learning and *Q*-learning with eligibility trace $Q(\lambda)$, are applied to solve the problem of optimal dynamic pricing of perishable products when demand is stochastic with unknown characteristics. The contribution of this paper is to propose a tested computational method (i.e., reinforcement learning) to solve the revenue management problem when information is incomplete and demand is non-stationary. In this paper we use two popular methods, *Q*-learning [16] and the *Q*-learning with eligibility traces, originally proposed by Peng and Williams [17].

[*] Corresponding author. Tel.: +44 (0) 1509 222721.
E-mail addresses: r.rana@lboro.ac.uk (R. Rana), Oliveira@essec.edu (F.S. Oliveira).

0305-0483/$ - see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.omega.2013.10.004

When selling products or services with well known statistical distributions the information about one product can be used to update the demand for similar products. For example, one may consider all flights for a particular origin–destination pair and a specific departure time each week. Since booking can start half a year in advance, or even earlier, this provides simultaneous learning opportunities for 26 or more concurrent episodes, shifted by 1 week relatively to each other. This can similarly be applied to hotel rooms, cabins on cruise liners, and cars at rental agencies. In all these services there are opportunities to book in advance, they are perishable, and there is a window for learning from the same service, from one time period to the next. For example, when pricing hotel rooms, the behaviour of demand for a given room type, on a Monday, can be used to price for the same room type the following Monday.

However, when a new product is launched, the demand patterns may be very different from past ones. In this case reinforcement learning will be even more helpful. If iPad 2 is launched and the demand profile is different from iPad, for the same time period, this difference is used by the algorithm to update the expectations about future demand for the product, implicitly, without having to explicitly compute a demand forecast.

The paper is organized as follows. Section 2 presents a literature review and places the contribution of this paper in comparison to the literature. Section 3 discusses how the model is formulated and describes how reinforcement learning is used to solve the dynamic pricing problem. Section 4 presents the analytical results, showing that *Q*-learning with eligibility converges to the optimal policy, and that the rate of learning is faster than with the simple *Q*-learning. Section 5 provides numerical results, and qualitative insights into the advantages of model-free reinforcement learning and compares the Q and $Q(\lambda)$ learning algorithms.

2. Related literature

The two main research areas relevant to this study are dynamic pricing and reinforcement learning, each of which is addressed in turn, together with a discussion of the contribution in this paper, where appropriate.

The majority of papers that address the problem of demand learning or estimation of pricing in the context of a single product, do so by assuming that one or more of the demand parameters are unknown. Recent examples include [1,3,18,19], who have incorporated real-time demand information in their models. Anjos et al. [18] develop a general methodology for implementing a pricing policy, and describe how the policy can be updated in real-time to react to changes in the predicted purchase patterns of consumers. Lin [1] forecast customer arrival rates in real-time using Bayesian statistics. Aviv and Pazgal [19] and Lin [1] assumed a known reservation price distribution. Berk et al. [20] investigate pricing of perishable products in menu costs recognizing that, although demand autocorrelation within the selling is important, it is often ignored in revenue management literature. They highlight that the complexity of modelling demand in an environment in which autocorrelation is a concern. In this case the *Q*-learning with eligibility traces algorithm has the merit of learning the pricing policy in a model free-environment and hence has the ability to implicitly incorporate autocorrelation of demand information within its policy. Zhao et al. [21] study a dynamic pricing problem for perishable goods to consumers who may exhibit inertia. They formulate this problem using the finite-horizon dynamic programming approach and derive an optimal dynamic pricing policy. Daso and Tong [22] have modelled the pricing of perishable products considering strategic buyers. Banerjee and Turner [23] have developed a model for pricing perishable goods based on differential equations, which is able to deal with group arrivals and continuous prices. Li et al. [24] have modelled dynamic pricing of perishable products with stochastic demand which they represented using randomness and fuzziness.

In most studies, there a few underlying assumptions in the model-based approaches, as in the case of the Bayesian learning. In this case, when the conditions of the model are violated, the policy is non-optimal. Lim and Shanthikumar [25] presented an approach for the single product dynamic pricing problem that accounts for errors in the underlying model at the optimization stage. They emphasized the importance of the underlying assumptions about the demand-rate when computing the optimal pricing policies. Besbes and Zeevi [26] developed nonparametric approaches that learn demand in a model free environment. However, there is a major issue with nonparametric approaches; the loss of tractability.

The reinforcement algorithms used in this paper produce look-up tables of value functions of all state-action pairs where the state is characterized as the capacity level and time until expiration and the action is the price. The value functions of a state-action pair are calculated by the immediate revenue gained and the expected future revenues. Using these algorithms allows the initialization of the value functions as the best estimated demand function and then observe the real-time demand and update the value functions. The decision-maker is then better informed than an agent that starts pricing assumes no prior knowledge. This is illustrated in the numerical experiments in Section 5.

There is limited literature in the area of revenue management using reinforcement learning to find an optimal pricing policy. Gosavi et al. [27] used reinforcement learning to develop a strategy for seating allocation and overbooking in order to maximize the average revenue gained by an airline. In particular, Raju et al. [28] used a reinforcement learning (*Q*-learning) algorithm to price products dynamically with customer segmentation. They considered an infinite horizon learning problem where there is no deadline for the sale of stock, and price changes according to queue length and time. Carvalho and Puterman [29] have also investigated dynamic pricing and reinforcement learning by studying maximization and learning problems in finite horizons for unlimited product quantities. They have focused on specific parametric forms of customer arrival distribution and on the probability of sales; the parameters are assumed to be fixed and unknown.

Cheng [30] applied the *Q*-learning approach to dynamic pricing in e-retailing. Cheng acknowledges that Lin's [1] approach to adjust price in response to changes in demand may not be plausible due to the computational complexity of using dynamic programming. Cheng has characterized demand in the same way as Lin except that the parameters of the model are learned using reinforcement learning. Price updates are made in real-time as the *Q*-learning algorithm produces a look-up table and, therefore, value function updates can be made with ease. Cheng focused on the computational advantages of using reinforcement learning and was not concerned with the accurate representation of demand.

This paper differentiates itself from the existing literature in dynamic pricing as it uses reinforcement learning in the context of pricing perishable products with non-stationary selling demand.

3. Model formulation of the dynamic pricing problem

3.1. Markov decision process

In this paper, the dynamic pricing problem of a perishable service is modeled as a discrete finite horizon Markov decision process (MDP). The dynamic pricing problem is formulated as a MDP because pricing is a real-time decision-making problem in a