Automatica 95 (2018) 385–398

Contents lists available at ScienceDirect

Automatica

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

Multiple stopping time POMDPs: Structural results & application in interactive advertising on social media☆

Vikram Krishnamurthy [a,*], Anup Aprem [b], Sujay Bhatt [a]

[a] Department of Electrical & Computer Engineering and Cornell Tech, Cornell University, NY, United States
[b] University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

Article history:
Received 27 June 2017
Received in revised form 18 March 2018
Accepted 19 May 2018

Keywords:
Partially observed Markov decision process
Multiple stopping time problem
Structural result
Monotone policies
Stochastic approximation
Monotone likelihood ratio dominance
Submodularity
Live social media
Scheduling
Interactive advertisement

ABSTRACT

This paper considers a multiple stopping time problem for a Markov chain observed in noise, where a decision maker chooses at most L stopping times to maximize a cumulative objective. We formulate the problem as a Partially Observed Markov Decision Process (POMDP) and derive structural results for the optimal multiple stopping policy. The main results are as follows: (i) The optimal multiple stopping policy is shown to be characterized by threshold curves Γ_l, for $l = 1, \ldots, L$, in the unit simplex of Bayesian Posteriors. (ii) The stopping sets S^l (defined by the threshold curves Γ_l) are shown to exhibit the following nested structure $S^{l-1} \subset S^l$. (iii) The optimal cumulative reward is shown to be monotone with respect to the copositive ordering of the transition matrix. (iv) A stochastic gradient algorithm is provided for estimating linear threshold policies by exploiting the structural results. These linear threshold policies approximate the threshold curves Γ_l, and share the monotone structure of the optimal multiple stopping policy. (v) Application of the multiple stopping framework to interactively schedule advertisements in live online social media. It is shown that advertisement scheduling using multiple stopping performs significantly better than currently used methods.

1. Introduction

Classical optimal stopping time problems are concerned with choosing a single time to take a stop action by observing a sequence of random variables in order to maximize a reward function. It has applications in numerous fields ranging from hypothesis testing (Lai, 1997), parameter estimation, machine replacement, multi-armed bandits and quickest change detection (Krishnamurthy, 2011; Krishnamurthy & Bhatt, 2016; Poor & Hadjiliadis, 2008). The optimal multiple stopping time problem generalizes the classical single stopping problem; the objective is to stop L-times to maximize the cumulative reward.

In this paper, motivated by the problem of interactive advertisement (ad) scheduling in personalized live social media, we consider a *multiple stopping time problem* in a partially observed Markov chain. Fig. 1 shows the schematic setup of the ad scheduling problem considered in this paper. The broadcaster (decision maker) in Fig. 1 wishes to schedule *at most* L ads to maximize the cumulative advertisement revenue.

Main results and organization. The multiple stopping time problem considered in this paper is a non-trivial generalization of the single stopping time problem, in that applying the single stopping policy multiple times does not yield the maximum possible cumulative reward; see Section 5 for a numerical example. Section 2 formulates the stochastic control problem faced by the decision maker (Broadcaster in Fig. 1) as a multiple stopping time partially observed Markov decision process (POMDP); the POMDP formulation is natural in the context of a partially observed multi-state Markov chain with multiple actions (L stops, continue). It is well known that for a POMDP, the computation of the optimal policy is PSPACE-complete (Krishnamurthy, 2016). Hence, we provide structural results on the optimal multiple stopping policy. The structural results are obtained by imposing sufficient conditions on the model — the main tools used are submodularity and stochastic dominance on the belief space of posterior distributions.

This paper has the following main results:

1. *Optimality of threshold policies:* Section 3.3 provides the main structural result of the paper. Specifically, Theorem 1 asserts that the optimal policy is characterized by up to L threshold curves,

☆ This research was funded by U. S. Army Research Office under grant 12346080, National Science Foundation under grant 1714180 and U.S. Air Force Office of Scientific Research under grant FA9550-18-1-0007. The material in this paper was partially presented at the 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), September 27–30, 2016, Monticello, IL, USA. This paper was recommended for publication in revised form by Associate Editor Hyeong Soo Chang under the direction of Editor Ian R. Petersen.

* Corresponding author.
E-mail addresses: vikramk@cornell.edu (V. Krishnamurthy), aaprem@ece.ubc.ca (A. Aprem), sh2376@cornell.edu (S. Bhatt).

https://doi.org/10.1016/j.automatica.2018.06.013
0005-1098/

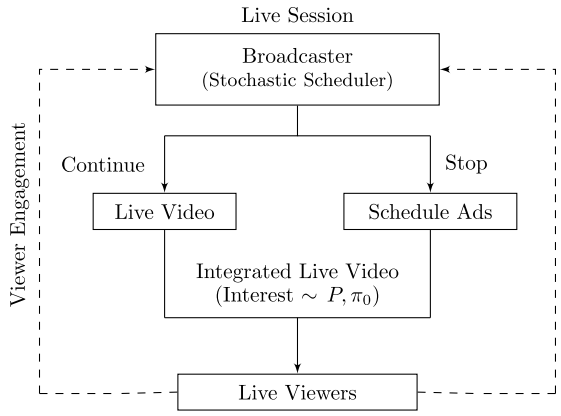

Fig. 1. Block diagram showing the stochastic scheduling problem faced by the decision maker (broadcaster) in advertisement scheduling on live media. The setup is detailed in Section 5 of the paper. The broadcaster wishes to schedule at most L-ads during the live session. To maximize advertisement revenue, the ads need to be scheduled when the interest in the content is high. The interest in the content cannot be measured directly, but noisy observations of the interest are obtained from the viewer engagement (viewer comments and likes) during the live session.

Γ_l on the unit simplex of Bayesian posteriors (belief states). To prove this result we use the monotone likelihood ratio (MLR) stochastic order since it is preserved under conditional expectations. However, determining the optimal policy is non-trivial since the policy can only be characterized on a partially ordered set (more generally, a lattice) within the unit simplex. We modify the MLR stochastic order to operate on line segments within the unit simplex of posterior distributions. Such line segments form chains (totally ordered subsets of a partially ordered set) and permit us to prove that the optimal decision policy has a threshold structure. In addition, similar to Nakai (1985), we show that the stopping sets (set of belief states at which the decision maker stops) have a nested structure.

2. *Monotonicity of cumulative reward with transition matrix:* Section 3.4 characterizes how the cumulative reward changes with respect to copositive ordering of the transition matrix. Specifically, Theorem 2 asserts that the optimal cumulative reward is monotone with respect to the copositive ordering of the transition matrix. The result can be used to implement reduced complexity posterior calculations for Markov chains with large dimension state space.

3. *Optimal Linear Threshold and their Estimation:* For the threshold curves $\Gamma_l, l = 1, \ldots, L$, Theorems 3 and 4 give necessary and sufficient conditions for the optimal linear hyperplane approximation (linear threshold policies) that preserves the structure of the optimal multiple stopping policy. Section 4 presents a simulation based stochastic gradient algorithm (Algorithm 1) to compute the optimal linear threshold policies. The advantage of the simulation based algorithm is that it is very easy to implement and is computationally efficient.

4. *Application to Interactive Advertising in live social media:* To illustrate the usefulness of the structural results for the multiple stopping time problem, we consider the application of interactive advertisement scheduling in personalized live social media. The problem of optimal scheduling of ads has been studied in the context of advertising in television; see Popescu and Crama (2015) and the references therein. However, scheduling ads on live online social media is different from scheduling ads on television in two significant ways (Kang & McAllister, 2011): (i) real-time measurement of viewer engagement (comments and likes on the content). The viewer engagement provides a noisy measurement of the underlying interest in the content. (ii) revenue is based on viewer engagement with the ads rather than a pre-negotiated contract. Section 5 uses a real dataset from Periscope, a popular personalized live streaming application owned by Twitter, to optimally schedule multiple ads ($L > 1$) in a sequential manner to maximize the advertising revenue.

Context and related literature. The problem of optimal multiple stopping has been well studied in the literature. In the classic L-secretary problem, independent and identically (i.i.d) observations are presented sequentially to the decision maker and the objective is to select L observations so as to maximize the sum of reward (a function of observation). The classical setting with i.i.d observations have been extended to consider variety of scenarios such as the observation times arising out of Poisson process (Stadje, 1987), observations with a joint distribution and possibly depending on the stopping times in Nikolaev (1999) and for random horizon in Krasnosielska-Kobos (2015). However, few works consider optimal multiple stopping over a partially observed Markov chain. The closest work is due to Nakai (1985) who considers optimal L-stopping over a finite horizon of length N in a partially observed Markov chain. In Nakai (1985), properties of the value function and the nested property of the stopping regions are derived. However, Nakai (1985) does not present an algorithm to compute the optimal policy utilizing the structural results. In addition, for many practical applications such as the interactive advertisement scheduling problem considered in this paper, the length of the horizon is not known apriori. Hence, this paper considers the multiple stopping problem over an infinite horizon, derives additional structural results compared to Nakai (1985) and provides a stochastic gradient algorithm to compute optimal approximation policies satisfying the structural results.

The optimal multiple stopping time problem can be contrasted to the recent work on sequential sampling with "causality constraints". Bayraktar and Kravitz (2015) considers the case where a decision maker is limited to a finite number of observations (sampling constraints) and must adaptively decide the observation strategy so as to perform quickest detection on a data stream. The extension to the case where the sampling constraints are replenished randomly is considered in Geng, Bayraktar, and Lai (2014). In the multiple stopping time problem, considered in this paper, there is no constraint on the observations and the objective is to stop at most L times to maximize the cumulative reward.

The optimal multiple stopping time problem, considered in this paper, is similar to the sequential scheduling problem with uncertainty (Alexander & Nikolaev, 2010) and the optimal search problem considered in the literature. Lobel, Patel, Vulcano, and Zhang (2015) considers the problem of finding the optimal launch times for a firm under strategic consumers and competition from other firms to maximize profit. However, in this paper, we deal with sequential scheduling in a partially observed case. The multiple-stopping problem considered in this paper is equivalent to a search problem where the underlying process is evolving (Markovian) and the searcher needs to optimally stop $L > 1$ times to achieve a specific objective.

Apart from interactive advertising, other applications of the multiple stopping problem include American options with multiple exercise times (Carmona & Touzi, 2008), L-commodities problem (Stadje, 1987), and investment decision making (Dahlgren & Leung, 2015).

2. Sequential multiple stopping and stochastic dynamic programming

In this section, we formulate the optimal multiple stopping time problem as a POMDP. In Section 2.3, we present a solution to the POMDP using stochastic dynamic programming. This sets the stage for Section 3 where we analyze the structure of the optimal policy.