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/

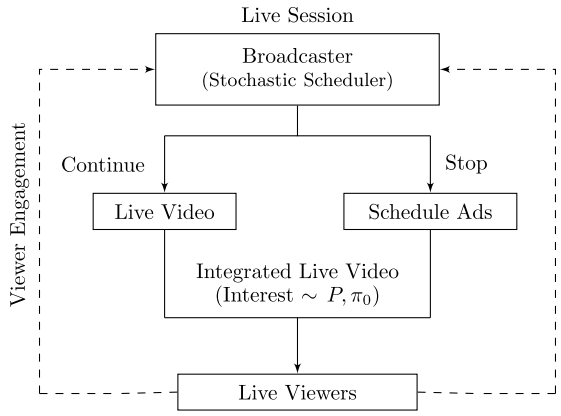

Fig. 1. Block diagram showing the stochastic scheduling problem faced by the decision maker (broadcaster) in advertisement scheduling on live media. The setup is detailed in Section 5 of the paper. The broadcaster wishes to schedule at most L-ads during the live session. To maximize advertisement revenue, 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.

2.1. Optimal multiple stopping: POMDP formulation

Consider a discrete time Markov chain X_t with state-space $\mathcal{S} = \{1, 2, \ldots, S\}$. Here, $t = 0, 1, \ldots$ denote discrete time. The decision maker receives a noisy observation Y_t of the state X_t at each time t. The decision maker wishes to stop at most L times over an infinite horizon. The positive integer L, is chosen a priori. At each time the decision maker either stops or continues, and obtains a reward that depends on the current state of the Markov chain. The objective of the decision maker is to opportunistically select the best time instants to stop so as to maximize the cumulative reward. This problem of stopping at most L times sequentially so as to maximize the cumulative reward corresponds to a multiple stopping time problem with L-stops.

The multiple stopping time problem consists of the following components:

1. *State Dynamics:* The Markov chain has time invariant transition matrix P and initial probability vector π_0; so

$$P(i, j) = \mathbb{P}(X_{t+1} = j | X_t = i), \ \pi_0(i) = \mathbb{P}(X_0 = i). \tag{1}$$

2. *Observations:* At each time instant t, the decision maker receives noisy observation Y_t of the state X_t. Denote, the conditional probability of receiving observation $y \in \mathcal{Y}$ $(Y_t = y)$ in state i $(X_t = i)$ by $B(i, y)$. Then, the time invariant observation distribution is

$$B(i, y) = \mathbb{P}\left(Y_t = y | X_t = i\right) \ \forall i \in \mathcal{S}, y \in \mathcal{Y}. \tag{2}$$

3. *Actions:* At each time instant t, the decision maker chooses an action $u_t \in \mathcal{A} = \{1\,(\text{Stop}), 2\,(\text{Continue})\}$ to either stop or to continue.

4. *Reward:* Choosing the stop action at time t, when there are l additional stops remaining, the decision maker accrues a reward[1] $r_l(X_t, a = 1)$, where X_t is the state of the Markov chain at time t. Similarly, if the decision maker chooses to continue, it will accrue $r_l(X_t, a = 2)$.

5. *Scheduling Policy:* The *history* available to the decision maker at time t is $Z_t = \{\pi_0, u_0, Y_1, \ldots, u_{t-1}, Y_t\}$. The scheduling policy μ, at each time t, maps Z_t to action u_t i.e. the action chosen at time t is $u_t = \mu(Z_t)$. Let $\mathcal{U}$ denote the set of admissible policies.

Objective: For $l \in \{1, 2, \ldots, L\}$, let τ_l denote the stopping time when there are l stops remaining, i.e.

$$\tau_l = \inf\{t : t > \tau_{l+1}, u_t = 1\}, \text{ with } \tau_{L+1} = 0. \tag{3}$$

For policy μ and initial belief π_0, the cumulative reward is:

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau_L - 1} \rho^t r_L(X_t, 2) + \rho^{\tau_L} r_L(X_{\tau_L}, 1) \right. \tag{4}$$
$$\left. + \sum_{t=\tau_L+1}^{\tau_{L-1}-1} \rho^t r_{L-1}(X_t, 2) + \cdots + \rho^{\tau_1} r_1(X_{\tau_1}, 1) \,\Big|\, \pi_0 \right\},$$

where the expectation is over the state dynamics and the observation distribution. In (4), $\rho \in [0, 1]$ denotes a user-defined economic discount factor.[2] Choosing $\rho < 1$ de-emphasizes the effect of decisions taken at later time instants on the cumulative reward.

[1] In interactive advertisement scheduling, the reward is indexed by the number of stops remaining to denote the varying ad revenue from the different ads placed during a session.

[2] In the multiple stopping time problem, considered here, $\rho = 1$ is allowed. For undiscounted problem ($\rho = 1$), the stopping times may not be finite and the objective in (4) becomes unbounded. However, the multiple stopping time problem considered in this paper will terminate in finite time: Assume $\bar{R} = \max_{i,l} r_l(i, 1) > 0$ i.e. the maximum stop reward is positive and $\underline{R} = \min_{i,l} r_l(i, 2) < 0$, i.e. the minimum reward to continue is negative. Then, it is clear that any optimal policy will stop in less than $\bar{T} = L\bar{R}/|\underline{R}|$ time steps. The intuition is that if $T > \bar{T}$ then the accumulated reward is negative and can be strictly improved by taking a stop action before $\bar{T}$.

The decision maker aims to compute the optimal strategy μ^* to maximize (4), i.e.

$$\mu^* = \arg\max_{\mu \in \mathcal{U}} J_\mu(\pi_0). \tag{5}$$

Remark 1. The above formulation is an instance of a stopping time POMDP. This is seen as follows: the objective in (4) can be expressed as an infinite horizon criteria by augmenting a fictitious absorbing state-0 that has zero reward, i.e. $r_0(0, u) = 0 \ u \in \mathcal{A}$. When L stop actions are taken, the system transitions to state 0 and remains there indefinitely. Then (4) is equivalent to the following discounted infinite horizon criteria:

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau_L - 1} \rho^t r_L(X_t, 2) + \rho^{\tau_L} r_L(X_{\tau_L}, 1) \right.$$
$$\left. + \cdots + \rho^{\tau_1} r_1(X_{\tau_1}, 1) + \sum_{t=\tau_1+1}^{\infty} \rho^t r_0(0, 2) \,\Big|\, \pi_0 \right\},$$

where the last summation is zero.

2.2. Belief state formulation of the objective

Let Π denote the belief space of S-dimensional probability vectors. The belief space is the unit $S - 1$ dimensional simplex:

$$\Pi = \left\{ \pi : 0 \le \pi(i) \le 1, \sum_{i=1}^{S} \pi(i) = 1 \right\}. \tag{6}$$

The belief state at time t, denoted by $\pi_t \in \Pi$, is the posterior probability of X_t given the history Z_t. The belief state is a sufficient statistic of Z_t, and evolves according to the following Hidden Markov model (HMM) Bayesian update (Krishnamurthy, 2016):

$$\pi_{t+1} = T(\pi_t, Y_{t+1}), \quad \text{where}$$
$$T(\pi, y) = \frac{B_y P' \pi}{\sigma(\pi, y)}, \quad \sigma(\pi, y) = \mathbf{1}_S' B_y P' \pi, \tag{7}$$

where, $B_y = \text{diag}\,(B(1, y), \ldots, B(S, y))$ and $\mathbf{1}_S$ represents the S-dimensional vectors of ones.

Using the smoothing property of conditional expectations, the objective in (4) can be reformulated in terms of belief state as:

$$J_\mu(\pi_0) = \mathbb{E}_\mu \left\{ \sum_{t=0}^{\tau_L - 1} \rho^t r'_{2,L} \pi_t + \rho^{\tau_L} r'_{1,L} \pi_{\tau_L} \right. \tag{8}$$
$$\left. + \sum_{t=\tau_L+1}^{\tau_{L-1}-1} \rho^t r'_{2,L-1} \pi_t + \cdots + \rho^{\tau_1} r'_{1,1} \pi_{\tau_1} + \sum_{t=\tau_1+1}^{\infty} \rho^t r'_{2,0} \pi_t \,\Big|\, \pi_0 \right\},$$

where $r_{u,l} = [r_l(1, u), \ldots, r_l(S, u)]'$. For the stopping time problem (8), there exists a stationary optimal policy. Since the belief state is a sufficient statistic of Z_t, (5) is equivalent to computing the optimal stationary policy $\mu^* : \Pi \times [L] \to \mathcal{A}$, where $[L] = \{1, 2, \ldots, L\}$, as a function of belief and number of stops remaining to maximize (8).

2.3. Stochastic dynamic programming

Computing the optimal policy μ^* to maximize (5) or equivalently (8) involves solving multiple stopping Bellman's dynamic

programming equation

$$\begin{aligned}
\mu^*(\pi, l) &= \arg\max_{u\in\mathcal{A}} Q(\pi, l, u),\\
V(\pi, l) &= \max_{u\in\mathcal{A}} Q(\pi, l, u),\\
Q(\pi, l, 1) &= r'_{1,l}\pi + \rho \sum_{y\in\mathcal{Y}} V\left(T(\pi, y), l-1\right)\sigma(\pi, y),\\
Q(\pi, l, 2) &= r'_{2,l}\pi + \rho \sum_{y\in\mathcal{Y}} V\left(T(\pi, y), l\right)\sigma(\pi, y).
\end{aligned} \tag{9}$$

Discussion: In (9), $V(\pi, l)$ denotes the optimal value function at belief π when l stops are remaining, and is the expected accumulated reward induced by the optimal policy μ^*. The optimal value function is the fixed point solution of the set of Bellman equations in (9). The fixed point solution can be obtained using the value iteration algorithm (see Appendix B). $Q(\pi, l, u)$ is the expected accumulated reward starting at belief π when l stops remaining, and taking action u and then using the optimal policy μ^*. The Bellman equations can be explained as follows: When a stop action ($u = 1$) is taken, the decision maker obtains an instantaneous reward $r'_{1,l}\pi$ and the number of stops remaining reduce by 1. When the continue action is taken ($u = 2$), the decision maker obtains an instantaneous reward of $r'_{2,l}\pi$, and the number of stops remaining is unaffected. The belief evolves according to (7).

Since the state-space Π is a continuum, Bellman's equation (9) or the value iteration algorithm in Appendix B does not translate into a practical solution methodology as $V(\pi, l)$ needs to be evaluated at each $\pi \in \Pi$. This, in turn, renders the computation of the optimal policy $\mu^*(\pi, l)$ intractable.[3]

3. Optimal multiple stopping: structural results

In this section, we derive structural results for the optimal policy (9) of the multiple stopping time problem. In Section 3.3, we show that under reasonable conditions on the POMDP parameters, the optimal policy is a monotone policy. In addition, in Section 3.4, we show the monotone property of the cumulative reward.

3.1. Definitions

Define stopping set S^l (the set of belief states where Stop is the optimal action), when l stops are remaining as:

$$S^l = \left\{\pi : \mu^*(\pi, l) = 1\right\}. \tag{10}$$

Correspondingly, the continue set (the set of belief states where Continue is the optimal action) is defined as

$$C^l = \left\{\pi : \mu^*(\pi, l) = 2\right\}. \tag{11}$$

Let $W(\pi, l)$ be defined as

$$W(\pi, l) = V(\pi, l) - V(\pi, l-1). \tag{12}$$

The stopping and continue sets in terms of W defined in (12) is as follows:

$$\begin{aligned}
S^l &= \{\pi | r'_l\pi \geq \rho \sum_y W(T(\pi, y), l)\sigma(\pi, y)\},\\
C^l &= \{\pi | r'_l\pi < \rho \sum_y W(T(\pi, y), l)\sigma(\pi, y)\}.
\end{aligned} \tag{13}$$

where, $r_l \triangleq r_{1,l} - r_{2,l}$.

Remark 2. For notational convenience, in this paper, without loss of generality, assume $r_{1,l} = r_l$ and $r_{2,l} = 0$. So, the decision maker accrues no reward for the continue action. Similarly, we consider $r_1 = r_2 = \cdots = r_L = r$, i.e. the rewards are not dependent on l.

In general, the stopping and continue sets can be arbitrary partitions of the simplex Π. However, in Section 3.3, we give sufficient conditions on the model so that these sets can be characterized by threshold curves. The question of computing the optimal policy, then, reduces to estimating the threshold curves.

It is worth pointing out that in the classical stopping POMDPs in Krishnamurthy (2016) with a single stop action, the stopping and continue sets are characterized in terms of convex value function. The key difficulty of the multiple stopping problem, considered in this paper, is that W being the difference of two convex value functions does not share the convex properties of the value function.

3.2. Assumptions

The main result below, namely, Theorem 1, requires the following assumptions on the reward vector, r (refer to Remark 3), the transition matrix, P and the observation distribution, B.

(A1) P is totally positive of order 2 (TP2), i.e. all second order minors are non-negative (see Definition 4 in Appendix A.1).
(A2) B is TP2.
(A3) The vector, $\bar{r} = (I-\rho P)r$, has decreasing elements, i.e. $\bar{r}(1) \geq \cdots \geq \bar{r}(S)$.

Discussion of Assumptions: Refer to Krishnamurthy (2016) for detailed discussions and examples of (A1)–(A3).

When $S = 2$, (A1) is valid when $P(1, 1) \geq P(2, 1)$. When $S > 2$, consider the tridiagonal transition matrix.[4] with $P(i, j) = 0, i > j + 2$ and $i < j - 2$. (A1) is valid if $P(i, i)P(i + 1, i + 1) \geq P(i + 1, i)P(i, i + 1)$.

(A2) holds for numerous examples. Examples include binomial, Poisson, geometric, Gaussian, exponential, etc. Table 1.1[5] and Table 1.2[5] in Müller and Stoyan (2002) contains a detailed list. In the numerical results in Section 5, we use the Poisson distribution where $B(i, y) = \frac{g_i^y \exp(-g_i)}{y!}$, where g_i is the mean of the Poisson distribution. (A2) is satisfied if g_i decreases monotonically with i. For a continuous observation distribution such as Gaussian whose mean is dependent on the state of the Markov chain (variance is fixed), (A2) is satisfied when the mean monotonically decreases with i.

(A3) is a joint condition on the reward vector and the transition matrix. Proposition 1, below, shows that (A3) and (A1) jointly imply that the reward vector r has decreasing elements. When $S = 2$, it can be verified that r having decreasing elements is sufficient for $(I - \rho P)\, r$ to have decreasing elements. For $S > 2$, (A3) is a stronger condition than having the elements of r decreasing.

(A3) is easy to interpret when P has additional structure. For example, consider a slowly varying Markov chain with $P = I + \epsilon Q$, where $Q(i, j) > 0, i \neq j$, $\sum_j Q(i, j) = 0$, and $\epsilon > 0$. Here $\frac{1}{\epsilon} > \max_i \sum_j |Q(i, j)|$ for P to be a valid transition matrix. Then (A3) is equivalent to r having decreasing elements. The reward vector r captures the preference of the decision maker — the highest reward is accrued in State 1.

Proposition 1 (*Proof in Appendix D.1*). *If P is TP2 and $(I - \rho P)r$ has decreasing elements, then r has decreasing elements.*

[3] It is well known that a finite horizon POMDP with finite observation space can be solved exactly, indeed the value function is piecewise linear and convex (Krishnamurthy, 2016). However, the problem is PSPACE complete; the worst case computational cost increases exponentially with the number of actions and doubly exponential with the time index.

[4] The transition matrices computed on real dataset in Section 5 follow a tridiagonal structure; refer to (24).

[5] Continuous distributions that satisfy (A2): Exponential, Normal, Gamma, Weibull, Lognormal, Beta. Discrete distributions that satisfy (A2): Poisson, Binomial, Geometric.

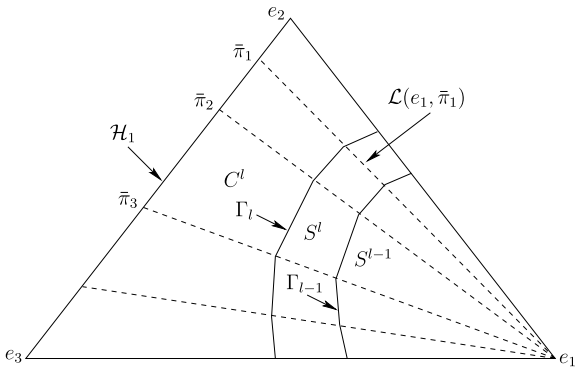

Fig. 2. Visual illustration of Theorem 1 when $S = 3$ so that the belief space is a 2-dimensional unit simplex (equilateral triangle). Each of the stopping sets S^l is characterized by a threshold curve Γ_l. Each of the threshold curve Γ_l intersects the line $\mathcal{L}(e_1, \bar{\pi})$ at most once.

3.3. Main result 1: optimality of threshold policies

The main result below (Theorem 1) states that the optimal policy is monotone with respect to the belief state π. However, for a monotone policy to be well defined, we need to first define the ordering between two belief states. For $S = 2$, the belief $\pi = \begin{bmatrix}1-\pi(2) & \pi(2)\end{bmatrix}$can be completely ordered with respect to $\pi(2) \in [0, 1]$. However, for $S > 2$, comparing belief states requires using stochastic orders which are partial orders. We will use the monotone likelihood ratio (MLR) (see Definition 1 in Appendix A.1); it is ideal for partially observed control problems since it is preserved under conditional expectation (Bayesian update).

Under reasonable conditions, Theorem 1 asserts that the optimal policy $\mu^*(\pi)$ is monotonically decreasing in π with respect to the MLR order. However, despite this monotonicity, determining the optimal policy is nontrivial since the policy can only be characterized on a partially ordered set. The main innovation in Theorem 1 is to modify the MLR stochastic order to operate on *lines* $\mathcal{L}(e_1, \bar{\pi})$ and $\mathcal{L}(e_S, \bar{\pi})$ (see Appendix A) within the belief space. 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.

Theorem 1. *Assume* (A1)–(A3)*. Then,*

A *There exists an optimal policy $\mu^*(\pi, l)$ that is decreasing on lines $\mathcal{L}(e_1, \bar{\pi})$, and $\mathcal{L}(e_S, \bar{\pi})$ in the belief space Π for each l.*

B *There exists an optimal switching curve Γ_l, for each l, that partitions the belief space Π into two individually connected sets S^l and C^l,such that the optimal policy is*

$$\mu^*(\pi, l) = \begin{cases} 1 & \text{if } \pi \in S^l \\ 2 & \text{if } \pi \in C^l \end{cases} \tag{14}$$

C $S^{l-1} \subset S^l$, $l = 1, 2, \ldots, L$.

The proof of Theorem 1 is given in Appendix C.4.

Discussion: Theorem 1A asserts that the optimal policy is monotonically decreasing on the line $\mathcal{L}(e_1, \bar{\pi})$, as shown in Fig. 2. Hence, on each line $\mathcal{L}(e_1, \bar{\pi})$ there exists a threshold above (in MLR sense) which it is optimal to *Stop* and below which it is optimal to *Continue*. Theorem 1B asserts, for each l, the stopping and continue sets are connected. Hence, there exists a threshold curve, Γ_l, as shown in Fig. 2, obtained by joining the thresholds, from Theorem 1A, on each of the line $\mathcal{L}(e_1, \bar{\pi})$. Theorem 1C proves the nested structure of the stopping sets: The stopping set when $l-1$ stops are remaining is a subset of the stopping set when there are l stops remaining.

In addition, Proposition 2, below, shows that the stopping set enclosed by the threshold curve is a union of convex sets and hence, the threshold curve is continuous and differentiable almost everywhere.

Proposition 2. *The stopping set S^l is a finite union of convex sets. (Proof in Appendix D.2.)*

3.4. Main result 2: monotonicity of cumulative reward with transition matrix

Large transition matrices, common in real world applications, require large number of numerical computations to keep track of the belief dynamics in (7). Knowledge of the belief state is crucial to implement the optimal policy using a scheduler. One approach to deal with the computational bottleneck is to select a suitable transition matrix "close" to the true transition matrix such that the computation of the belief update is cheaper. It was shown in Krishnamurthy and Rojas (2014) that convex optimization techniques can be used to compute reduced rank matrices that bound (in terms of copositive ordering- Definition 6 in Appendix) the true transition matrix P from above and below, i.e. $\underline{P} \preceq P \preceq \bar{P}$. Computing the belief state in (7) requires $\mathcal{O}(S^2)$ computations, which could be expensive for large dimensional state space. The computational cost is reduced by using low rank (rank R) transition matrices ($\underline{P}$ and $\bar{P}$) which requires only $\mathcal{O}(RS)$ numerical operations.

This leads us to the following question: How does the optimal cumulative reward of a multiple stopping time problem vary with transition matrix P? The main result below shows that if the transition matrices are partially ordered with respect to the copositive ordering so that $P \succeq \bar{P}$ then $J_{\mu^*(P)} \geq J_{\mu^*(\bar{P})}$.

Theorem 2. *Consider two multiple stopping time problems with transition matrices P and $\bar{P}$, respectively, where $P \succeq \bar{P}$ with respect to copositive ordering (Definition 6 in Appendix). If (A1) to (A3) hold, then the optimal cumulative rewards satisfy $J_{\mu^*(P)} \geq J_{\mu^*(\bar{P})}$.*

The proof follows from Krishnamurthy (2016, Theorem 14.8.1).

Discussion: Theorem 2 asserts that larger transition matrix (with respect to the copositive order) always results in a larger optimal reward. This is useful in obtaining bounds on the achievable rewards in applications like interactive advertisement scheduling, where the interest dynamics change slowly over time. Also, the performance loss from using a low rank transition matrix for interest dynamics – to reduce the complexity of the real time scheduler – can be characterized.

Summary: This section derived the structural results of the optimal multiple stopping problem. The main structural result is in Theorem 1. Theorem 1 generalizes the results in Nakai (1985). In addition to the nested property in Nakai (1985), Theorem 1 characterizes the optimal policy by up to L threshold curves. Theorem 2 established the monotonicity of the optimal cumulative reward with respect to the copositive ordering of the transition matrix.

4. Stochastic gradient algorithm for estimating optimal linear threshold policies

In light of Theorem 1, computing the optimal policy reduces to estimating L-threshold curves in the unit simplex (belief space), one for each of the L-stops. The threshold curves can be approximated by any of the standard basis functions. In this paper, we will restrict the approximation to *linear* threshold policies, i.e. policies of the form given in (15). However, any such approximation needs to capture the essence of Theorem 1, i.e. the optimal policy is MLR decreasing on lines, connected and satisfy the nested property. We

call such linear threshold policies (that captures the essence of Theorem 1) as the *optimal* linear threshold policies.

Section 4.1 derives necessary and sufficient conditions to characterize such linear threshold policies. Algorithm 1 in Section 4.2 is a simulation based algorithm to compute the optimal linear threshold policies. The simulation based algorithm is computationally efficient (see comments at end of Section 4.2).

4.1. Structure of optimal linear threshold policies for multiple stopping

We define a linear parametrized policy on the belief space Π as follows. Let $\theta_l \in \mathbb{R}^{S-1}$ denote the parameters of linear hyperplane. Then, linear threshold policies as a function of the belief π and the number of stops remaining l, are defined as

$$\mu_\theta(\pi, l) = \begin{cases} 1 & \text{if } \begin{bmatrix} 0 & 1 & \theta_l \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} \le 0 \\ 2 & \text{otherwise .} \end{cases} \tag{15}$$

The linear policy $\mu_\theta(\pi, l)$ is indexed by θ to show the explicit dependence of the parameters on the policy. In (15), $\theta = (\theta_1, \theta_2, \ldots, \theta_L) \in \mathbb{R}^{L\times(S-1)}$ is the concatenation of the θ_l vectors, one for each of the L-stops.

Discussion: We will briefly discuss (15): Given a general linear policy of the form $\alpha'\pi \le \beta$, the specific form in (15) is obtained using (i) the sum constraint on the belief π, i.e. $\sum_{i=1}^{S}\pi(i) = 1$, (ii) Scale invariance: For any positive constant c, $\alpha'\pi \le \beta \Rightarrow c\alpha'\pi \le c\beta$. Also, notice that the dimension of both $\begin{bmatrix} 0 & 1 & \theta_l \end{bmatrix}$ and $\begin{bmatrix} \pi & -1 \end{bmatrix}$ is $S+1$, since $\theta_l \in \mathbb{R}^{S-1}$ and $\pi \in \mathbb{R}^S$.

In Theorem 1A, it was established that the optimal multiple stopping policy is MLR decreasing on specific lines within the belief space, i.e. for $\pi_1 \ge_{\mathcal{L}_i} \pi_2$, $\mu(\pi_1, l) \le \mu(\pi_2, l)$; $i = 1, S$. Theorem 3 gives necessary and sufficient conditions on the coefficient vector θ_l such that $\pi_1 \ge_{\mathcal{L}_i} \pi_2$, $\mu_\theta(\pi_1, l) \le \mu_\theta(\pi_2, l)$; $i = 1, S$.

Theorem 3. *A necessary and sufficient condition for the linear threshold policies $\mu_\theta(\pi, l)$ to be*

(1) *MLR decreasing on line $\mathcal{L}(e_1)$, iff $\theta_l(S-1) \ge 0$ and $\theta_l(i) \ge 0$, $i \le S-2$.*
(2) *MLR decreasing on line $\mathcal{L}(e_S)$, iff $\theta_l(S-1) \ge 0$, $\theta_l(S-2) \ge 1$ and $\theta_l(i) \le \theta_l(S-2)$, $i < S-2$.*

The proof of Theorem 3 is similar to Theorem 12.4.1 in Krishnamurthy (2016) and hence omitted. In Theorem 3, $\theta_l(i)$ denotes the ith element of $S-1$ dimensional vector θ_l.

Discussion: By Theorem 3, the constraints on the parameters θ ensure that only MLR decreasing linear threshold policies are considered; the necessity and sufficiency imply that non-monotone policies are not considered, and monotone policies are not left out.

Theorem 1B established that the optimal stopping sets are connected, which is satisfied trivially since we approximate the threshold curve using a linear hyperplane. Theorem 4 below provides sufficient conditions so that the parametrized linear threshold curves satisfy the nested property established in Theorem 1C.

Theorem 4 (*Proof in Appendix C.6*)**.** *A sufficient condition for the linear threshold policies in (15) to satisfy the nested structure in Theorem 1C is given for each l by*

$$\begin{aligned} &\theta_{l-1}(S-1) \le \theta_l(S-1) \\ &\theta_{l-1}(i) \ge \theta_l(i) \quad i < S-1, \end{aligned} \tag{16}$$

4.2. Simulation-based stochastic gradient algorithm for estimating linear threshold policies

We now estimate the optimal linear threshold policies using a simulation based stochastic gradient algorithm (Algorithm 1). The algorithm ensures that the estimated policies satisfy the conditions in Theorems 3 and 4.

The optimal policy of a multiple stopping time problem maximizes the expected cumulative reward J_μ in (4). In Algorithm 1, we approximate J_μ over a finite time horizon (N), as J_N which is computed as:

$$J_N(\theta) = \mathbb{E}_{\mu_\theta}\left\{ \sum_{l=1}^{L} \rho^{\tau_l} r'\pi_{\tau_l} \;\middle|\; \tau_l \le N; \forall l \right\}. \tag{17}$$

For the optimal policy μ^*, a horizon of length N and the discount factor of ρ, $|J_{\mu^*} - J_N|_2 \le \frac{\rho^N}{1-\rho}\max_{l,x,u}|r_l(x,u)|$ (Krishnamurthy, 2016 Theorem 7.6.3).[6]

Algorithm 1 Stochastic Gradient Algorithm for Optimal Multiple Stopping

Require: POMDP parameters satisfy (A1)–(A2).
1: Choose initial parameters ϕ_0 and initial linear threshold policies $\mu_{\theta^{\phi_0}}$ using (15).
2: **for** iterations $n = 0, 1, 2, \ldots$: **do**
3: Evaluate $J_N(\theta^{\phi_n + c_n\omega_n})$ and $J_N(\theta^{\phi_n - c_n\omega_n})$ using (17)
4: SPSA: Gradient estimate $\hat{\nabla}_\phi J_N(\theta^{\phi_n})$ using (19).
5: Update parameter vector ϕ_n to ϕ_{n+1} using (20).

Algorithm 1 is a stochastic gradient algorithm that generates a sequence of estimates θ_n, that converges to a local maximum. It requires the computation of the gradient: $\nabla_\theta J_N(\cdot)$. Evaluating the gradient in closed form is intractable due to the non-linear dependence of $J_N(\theta)$ on θ. We can estimate $\hat{\nabla}_\theta J_N(\cdot)$ using a simulation based gradient estimator. There are several such simulation based gradient estimators available in the literature including infinitesimal perturbation analysis, weak derivatives and likelihood ratio (score function) methods. For simplicity, we use the SPSA algorithm (Spall, 2005), which estimates the gradient using a finite difference method.

To make use of the SPSA algorithm, we convert the constrained optimization problem in θ (constraints imposed by Theorems 3 and 4) into an unconstrained problem using spherical co-ordinates as follows:

$$\theta_l^\phi(i) = \begin{cases} \phi_1^2(S-1)\displaystyle\prod_{\ell=l}^{L-1}\sin^2(\phi_\ell(S-1)) & i = S-1 \\ 1 + \phi_1^2(S-2)\displaystyle\prod_{\ell=2}^{l}\sin^2(\phi_\ell(S-2)) & i = S-2 \\ \theta_l(S-2)\displaystyle\prod_{\ell=1}^{L}\sin^2(\phi_\ell(i)) & i < S-2. \end{cases} \tag{18}$$

It can be verified that the parametrization, θ^ϕ in (18), satisfies the conditions in Theorems 3 and 4. For example, consider $i = S-1$, then the product term involving $\sin(\cdot)$ ensures that $\theta_{l-1}(S-1) \le \theta_l(S-1)$ (the first part of Theorem 4).

Following Spall (2005), the gradient estimate using SPSA is obtained by picking a random direction ω_n, at each iteration n. The estimate of the gradient is then given by

$$\hat{\nabla}_\phi J_N(\theta^{\phi_n}) = \frac{J_N(\theta^{\phi_n + c_n\omega_n}) - J_N(\theta^{\phi_n - c_n\omega_n})}{2c_n}\omega_n, \tag{19}$$

$$\text{where } \omega_n(i) = \begin{cases} -1 & \text{with probability 0.5} \\ +1 & \text{with probability 0.5.} \end{cases}$$

[6] Given an error tolerance ε, the required horizon can be calculated as $N > \log\left(\frac{(1-\rho)\varepsilon}{\max_{l,x,u}|r_l(x,u)|}\right)/\log\rho$.

The two $J_N(\cdot)$ terms in the numerator of (19) is estimated using the finite time horizon approximation (17). A more detailed description of the finite time horizon approximation in given in Algorithm 2 in Appendix E. Using the gradient estimate in (19), the parameter update is as follows (Spall, 2005):

$$\phi_{n+1} = \phi_n + a_n \hat{\nabla}_\phi J_N(\theta^{\phi_n}). \tag{20}$$

The parameters a_n and c_n are typically chosen as (Spall, 2005):

$$\begin{aligned} a_n &= \varepsilon(n+1+\varsigma)^{-\kappa} \quad 0.5 < \kappa \leq 1, \text{ and } \varepsilon, \varsigma > 0 \\ c_n &= \mu(n+1)^{-\upsilon} \qquad 0.5 < \upsilon \leq 1 \; \mu > 0 \end{aligned} \tag{21}$$

The decreasing step size stochastic gradient algorithm, Algorithm 1, converges to a local optimum with probability one. There are several methods available in the literature that can be used for stopping criteria in Step 2 of Algorithm 1 (Spall, 2005). In this paper, we used the following criteria: (i) Small gradient: $\|\hat{\nabla}_\phi J_N(\theta^{\phi_n})\|_2 \leq \varepsilon$. (ii) Max Iteration: Iterations are stopped when a maximum number is reached.

At each iteration of Algorithm 1, evaluating the gradient estimate in (19) requires two POMDP simulations. However, this is independent of the number of states, the number of observations or the number of stops.

5. Numerical examples: interactive advertising in live social media

This section has three parts. In Section 5.1, we illustrate the main result of the paper using numerical examples. Second, using a Periscope dataset, we study how the multiple stopping problem can be used to schedule advertisements in live social media. We show numerically that the linear threshold scheduling policies (derived in Section 4) outperforms conventional techniques for scheduling ads in live social media. Finally, we illustrate the performance of the linear threshold policies for a large size POMDP by comparing with the SARSOP algorithm, which is a popular suboptimal POMDP solver.

5.1. Synthetic data

This section has four parts. First, we visually illustrate the optimal multiple stopping policy, using numerical examples, for $S = 3$. The objective is to illustrate how the assumptions in Section 3.2 affect the optimal multiple stopping time policy. The optimal policy can be obtained by solving the dynamic programming equations in (9) and can be computed approximately by discretizing the belief space. The belief space Π, for all examples below, was uniformly quantized into 100 states, using a finite grid approximation. Second, we illustrate how the optimal accumulated rewards varies with the number of stops. Third, we benchmark the performance of linear threshold policies (obtained using Algorithm 1) against optimal multiple stopping policy. Finally, we illustrate the advantage of structural results for designing approximation algorithm by comparing the performance of the linear threshold policies in Section 4 against the popular softmax parametrization, which are not constrained to satisfy the structural results.

Example 1. *POMDP parameters:* Consider a Markov chain with 3−states with the transition matrix P and the reward vector specified in (22). The observation distribution is given by $B(i, y) = \frac{g_i^y \exp(-g_i)}{y!}$, i.e. the observation distribution is Poisson with state dependent mean vector g given in (22). It is easily verified that the transition matrix, the observation distribution and the reward vector satisfy the conditions (A1)–(A3).

$$P = \begin{bmatrix} 0.2 & 0.1 & 0.7 \\ 0.1 & 0.1 & 0.8 \\ 0 & 0.1 & 0.9 \end{bmatrix}, g = \begin{bmatrix} 12 & 7 & 2 \end{bmatrix}', r = \begin{bmatrix} 9 & 3 & 1 \end{bmatrix}' \tag{22}$$

We choose $L = 5$, i.e. the decision maker wishes to stop at most 5 times. Fig. 3 shows the stopping sets S^5 and S^1. It is evident from Fig. 3 that the optimal policy is monotone on lines, stopping sets are connected and satisfy the nested property; thereby illustrating Theorem 1.

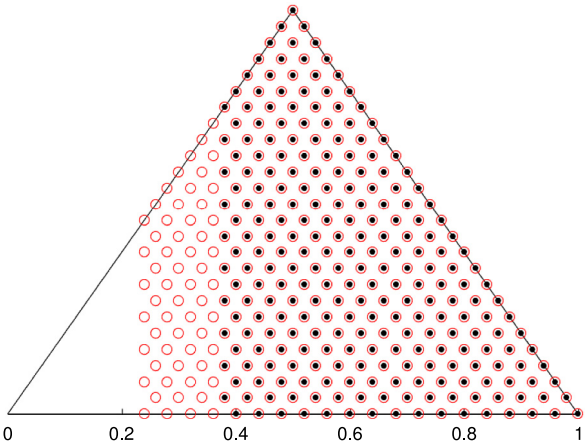

Fig. 3. Example 1: S^1 (shown in black) and S^5 (shown in red) obtained by solving the dynamic programming (9). The figure illustrates monotone, connected and the nested structure of the stopping sets ($S^{l-1} \subset S^l$), in Theorem 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

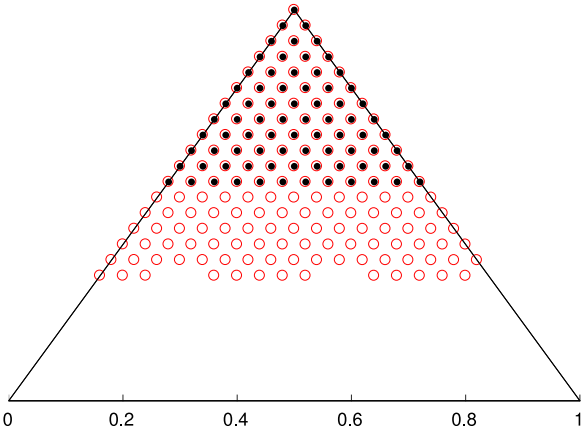

Fig. 4. Example 2: Optimal policy when (A3) is violated. S^1 is shown in black and S^5 is shown in red. The monotone property of Theorem 1A is violated.

Example 2. Consider the same parameters as in Example 1, except reward $r = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}'$ which violates (A3). Fig. 4 shows the optimal multiple stopping policy in terms of the stopping sets. As can be seen from Fig. 4 that the optimal policy does not satisfy the monotone property (Theorem 1A). However, the nested property continues to hold.

Example 3. Consider the same parameters as in Example 1, except $L = 2$ and $r_1 = \begin{bmatrix} 9 & 3 & 1 \end{bmatrix}'$and $r_2 = \begin{bmatrix} 3 & 9 & 1 \end{bmatrix}'$. (A3) is violated for $l = 2$. Fig. 5 shows the optimal multiple stopping policy in terms of the stopping sets. As can be seen from Fig. 5 that

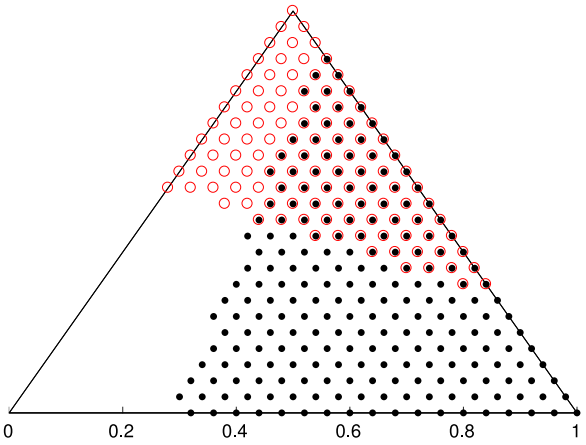

Fig. 5. Example 3: Optimal policy when (A3) is violated. S^1 is shown in black and S^2 is shown in red. The stopping sets are not nested.

Table 1
Optimal accumulated reward (normalized w.r.t $L = 1$) versus number of stops. As the number of stops increases the accumulated reward increases. The table was generated by solving the dynamic programming equations in (9). The accumulated reward is with a starting belief $\pi_0 = \left(\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3}\right)$.

L	1	2	3	4	5
reward	1	1.66	2.12	2.46	2.75

the optimal policy does not satisfy the monotone property or the nested property.

Thus, the conditions (A1)-(A3) of Theorem 1 are useful in the sense that when they are violated, there are examples where the optimal policy does not have the monotone or nested property.

Optimal accumulated reward against L: Consider Example 1 with POMDP parameters in (22). At each stop, we accumulate a reward. It is easy to see that as the number of stops increase, the reward accrued will also increase. Table 1 illustrates that this is indeed the case. The values in Table 1 were obtained by solving the dynamic programming equations in (9) for various values of L ranging from 1 to 5.

Performance of linear threshold policies: In order to benchmark the performance of optimal linear threshold policies (that satisfy the constraints in Theorems 3 and 4), we ran Algorithm 1 for Example 1 (parameters in (22)). The performance was compared based on the expected cumulative reward between the optimal policy and the linear threshold policies for 1000 independent runs. The following parameters were chosen for the SPSA algorithm $\mu = 2, \upsilon = 0.2$, $\varsigma = 0.5, \kappa = 0.602$ and $\varepsilon = 0.1667$; these values are as suggested in Spall (2005). It was observed that there is a 12% drop in performance of the linear threshold policies compared to the optimal multiple stopping policy.

Advantage of parametrization satisfying structural results: Here, we illustrate the advantage of parametrization of the policy to satisfy the structural results in Theorem 1. The softmax function is a popular parametrization for decision-making and is widely used in reinforcement learning (Sutton & Barto, 1998). Consider the following softmax parametrization of the policy

$$\Pr(\mu(\pi, l) = u) = \frac{\exp\left(\begin{bmatrix}0 & \theta_{l,u}\end{bmatrix}'\pi\right)}{\sum_{u=1}^{2}\exp\left(\begin{bmatrix}0 & \theta_{l,u}\end{bmatrix}'\pi\right)}. \tag{23}$$

In (23), $\Pr(\mu(\pi, l) = u)$ denotes the probability of taking action u (either 'Stop' or 'Continue') as a function of belief π and number of stops remaining l. The parameters in (23) $\theta_{l,u} \in \mathbb{R}^{S-1}; l = 1, \ldots, L\ \ u = 1, 2$ are indexed by number of stops remaining and the actions. Compared to linear threshold policies in (15), the policies in (23) are not restricted to satisfy the structural results in Theorem 1. Algorithm 3 in Appendix E summarizes the computation of the finite time horizon approximation with the softmax parametrization in (23).

Comparing the expected cumulative reward, we find that the optimal policy and the linear threshold policies outperform the softmax parametrization by 40% and 30%, respectively. Hence, this illustrates the advantage of taking into account the structure of the optimal policy while designing algorithms for computing an approximation policy.

5.2. Real dataset: interactive ad scheduling on periscope using viewer engagement

We now formulate the problem of interactive ad scheduling on *live* online social media as a multiple stopping problem and illustrate the performance of linear threshold policies using a Periscope dataset.[7] Periscope is a popular live personalized video streaming application where a broadcaster interacts with the viewers via live videos. Each such interaction lasts between $10 - 20$ minutes and consists of: (i) A broadcaster who starts a live video using a handheld device. (ii) Video viewers who engage with the live video through comments and likes.

Dataset: The dataset in Wang et al. (2016) contains details of all public broadcasts on the Periscope application from May 15, 2015 to August 20, 2015. The dataset consists of timestamped events: time instants at which the live video started/ended; time instants at which viewers join; and, time instants at which the viewers engage using *likes* and *comments*. In this paper, we consider viewer engagement through *likes*, since *comments* are restricted to the first 100 viewers in the Periscope application.

Ad scheduling Model

Here we describe how the model in Section 2 can be adapted to the problem of interactive ad scheduling in live video streaming; see Fig. 1 for the setup.

1. *Interest Dynamics:* In live online social media, it is well known that the viewer engagement is correlated with the interest of the content being streamed or broadcast. Markov models have been used to model interest in online games (Baldominos, Esperanza, Marrero, & Saez, 2016), and in online social networks (Benevenuto, Rodrigues, Cha, & Almeida, 2009). We therefore model the interest in live video as a Markov chain, X_t, where the different states denote the level of interest in the live content. The states are ordered in the decreasing order of interest.

Homogeneous Assumption: Periscope utilizes the Twitter network to link broadcasters with the viewers and hence shares many of the properties of the Twitter social network. Different sessions of a broadcaster, therefore, tend to follow similar statistics due to the effects of social selection and peer influence (Lewis, Gonzalez, & Kaufman, 2012). It was shown in Hamilton, Garretson, and Kerne (2014) that live sessions on live online gaming platforms can be viewed as *communities* and communities in online social media have similar information consumption patterns (Del Vicario, Bessi, Zollo, Petroni, Scala, Caldarelli, Stanley, & Quattrociocchi, 2016). We therefore model the interest dynamics as a time homogeneous Markov chain.

2. *Engagement Dynamics:* The interest in the video, X_t, cannot be measured directly by the broadcaster and has to be inferred from

[7] We use the dataset in Wang, Zhang, Wang, Zheng, and Zhao (2016), which can be downloaded from http://sandlab.cs.ucsb.edu/periscope/. Wang et al. (2016) deals with the performance of Periscope application in terms of delay and scalability.

the viewer engagement, denoted by Y_t. Since the viewer engagement measures the number of likes in a given time interval, we model it using a Markov modulated Poisson distribution. Denote the rate of the Poisson observation process when the interest is in state i by g_i. The observation probability in (2) can be obtained using $B(i, y) = g_i^y \exp(-g_i)/y!$.

3. *Broadcaster Revenue:* The ad revenue in online social media depends on the click rate (the probability that the ad will be clicked). In a recent research, Adobe Research[8] concluded that video viewers are more likely to engage with an ad if they are interested in the content of the video that the ad is inserted into. The reward vector in Section 2.1 should capture the positive correlation that exists between interest in the videos and the click rate (Lehmann, Lalmas, Yom-Tov, & Dupret, 2012). Since the information regarding the click rate and actual number of viewers are not available in the dataset, we choose the reward vector r to be a vector of decreasing elements, each being proportional to the reward in that state, such that (A3) is satisfied.

4. *Broadcaster operation:* The broadcaster wishes to schedule at most L ads at instants when the interest is high. Here, we choose.[9] the number of stops $L = 5$. At each discrete time, after receiving the observation Y_t, the broadcaster either stops and schedules an ad or continues with the live stream; see Fig. 1 The ad scheduling model that we consider in this paper assumes that the interest in the content does not change with scheduling ads. This is a simplified model when the live video content is paused to allow for advertisements, as in Twitch. However, the model captures the *in-video overlay* ads that are popular in YouTube Live. In video overlay ads, the advertisement is shown in a portion of the screen (typically below). Here, it is safe to assume that the interest is not affected by ad-scheduling.

5. *Broadcaster objective:* The objective of the broadcaster is given by (4). It aims to schedule ads when the content is interesting, so as to elicit maximum number of clicks, thereby maximizing the expected revenue. In personalized live streaming applications like Periscope, the discount factor in (4) captures the "impatience" of live broadcaster in scheduling ads.

The above model and formulation correspond to a multiple stopping problem with L stops, as discussed in Section 2. Theorem 1 establishes structural results on the optimal ad scheduling policy. In addition, personalized live social media applications like Periscope need to work seamlessly on a mobile smart phone platform, where computing resources are limited. Hence, advertising scheduling algorithm should have minimal real-time computational requirements. The real-time computational requirement for implementing any POMDP policy consists of (i) Updating the belief using the HMM Bayesian update in (7), (ii) Evaluating the policy using the updated belief to obtain a decision. The linear threshold policies reduce the computational requirement of obtaining a decision from an updated belief.

Below, we describe how to estimate the model parameters from the data (viewer engagement Y_t) for computing the linear threshold policies using Algorithm 1.

Estimation of parameters: The live video sessions in Periscope have a range of 10–20 min (Wang et al., 2016). The viewer engagement information consists of a time series of *likes* obtained by sampling the timestamped likes at a 2-s interval. Sampling at a 2-s interval, each session provides 1000 data points. The model parameters P and B are computed using maximum likelihood estimation. Since the interest dynamics are time homogeneous, we utilize data from multiple sessions to estimate the parameters P and B. The model was validated using the QQ-plot (see Fig. 6) of normal pseudo-residuals. The estimated value of the transition matrix P and the state dependent mean g of a popular live session are given as:

$$P = \begin{bmatrix} 0.733 & 0.266 & 0.000 & 0.000 \\ 0.081 & 0.718 & 0.201 & 0.000 \\ 0.000 & 0.214 & 0.670 & 0.116 \\ 0.000 & 0.000 & 0.222 & 0.778 \end{bmatrix}, g = \begin{bmatrix} 38 \\ 21 \\ 10 \\ 1 \end{bmatrix}. \quad (24)$$

The model order dimension was estimated using the penalized likelihood criterion; specifically Table 2 shows the model order selection using the Bayesian information criterion (BIC). The likelihood values in Table 2 were obtained using an Expectation–Maximization algorithm. In Table 2, $S = 4$ has the lowest BIC value. The reward vector was chosen as $r' = [4 \quad 3 \quad 2 \quad 1]$, and satisfies (A3) for $\rho \in [0, 1]$.

Table 2
BIC model order selection for the popular live session. The maximum likelihood estimated parameters are given in (24). The BIC criteria were run for S varying from 2 to 12 (only values for 2–6 are shown below). $S = 4$ has the lowest BIC value.

S	$-\log(\mathscr{L})$	BIC $= -2\log(\mathscr{L}) + n\log(N)$
2	−4707.254	9535.053
3	−4190.652	8601.122
4	−3969.955	**8287.364**
5	−3951.155	8405.764
6	−3887.453	8462.725

- $\mathscr{L}$ denotes the likelihood value.
- n denotes the number of parameters: $n = S^2 + S - 1$.
- N denotes the number of observations. Here, $N = 10^4$.

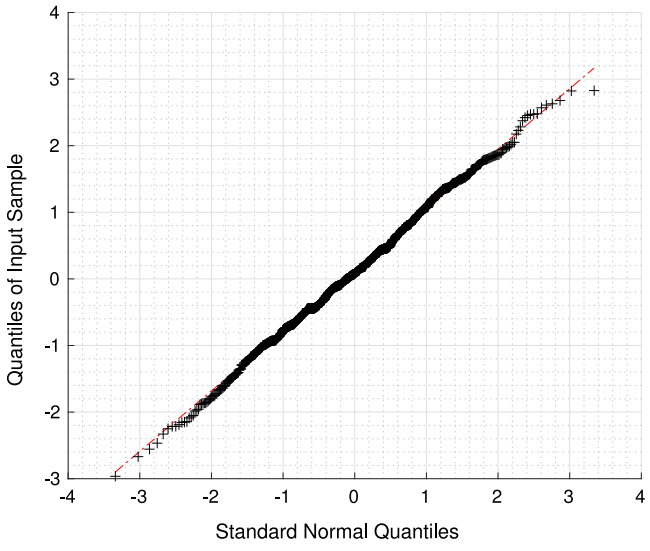

Fig. 6. The maximum likelihood estimated parameters are given in (24). The QQ-plot is used for validating the goodness of fit. The linearity of the points suggests that the estimated parameters in (24) are a good fit.

5.2.1. Multiple ad scheduling: performance results

We now compare the linear threshold scheduling policies (obtained from Algorithm 1) with two existing schemes:

(1) *Periodic:* Here, the broadcaster stops periodically to advertise. Twitch,[10] for example, uses periodic ad scheduling (Smith, Obrist, & Wright, 2013). Periodic advertisement

[8] https://gigaom.com/2012/04/16/adobe-ad-research/.

[9] Most of the popular Periscope sessions last 15–30 min. Broadcast television usually average 13.5 mins per hour of advertisement or approximately one ad every 5 min. Hence, we choose the number of advertisements $L = 5$.

[10] Twitch is a video platform that focuses primarily on video gaming. In 2015, Twitch had more than 1.5 million broadcasters and 100 million visitors per month.

scheduling is also widely used for pre-recorded videos on social media platforms like YouTube.

(2) *Heuristic:* Here, the broadcaster solves a classical single stopping problem ($L = 1$ in Section 2). The single stopping policy is used to schedule all the advertisements.

Performance Results: It was seen that the optimal linear threshold policies outperforms conventional periodic scheduling by 25% and the heuristic scheduling by 10%. The periodic scheme performs poorly because it does not take into account the viewer engagement or the interest in the content while scheduling ads. The multiple stopping policy, in comparison to the heuristic scheme, takes into account the fact that L-ads need to be scheduled and hence, is optimal.

5.3. Large dimensional state space models & Comparison with SARSOP

We consider a Markov chain with 100 states. The transition matrix and observation distribution are generated as discussed in Krishnamurthy and Rojas (2014). In order for the transition matrix P satisfy the TP2 assumption in (A1), we use the following approach: First construct a 10-state transition matrix $A = \exp(Qt)$, where Q is a tridiagonal generator matrix (off-diagonal entries are non-negative and row sums to 0) and $t > 0$. Since Kronecker product preserves TP2 structure, we let $P = A \otimes A$. The observation distribution B, containing 100 observations satisfying (A2) is similarly generated. The reward vector is chosen as follows: $r = [100, 99, \ldots, 1]$. The number of stops is $L = 5$.

Because of the large state space dimension, computing the optimal policy using dynamic programming is intractable. We compare linear threshold policies (obtained through Algorithm 1), the heuristic policy and periodic policy (described in the Section 5.2), in terms of the expected cumulative reward by each of the policy. Also, we compare the linear threshold policy against the state-of-the-art solver for POMDP: SARSOP (an approximate POMDP planning algorithm) (Kurniawati, Hsu, & Lee, 2008).

Table 3 shows the normalized cumulative reward by each of the policies. The expected reward was calculated using 1000 independent Monte Carlo simulations. From Table 3 we observe that the linear threshold policy and heuristic policy outperforms periodic scheduling by a factor of 2. Also, the linear threshold policy outperforms the heuristic policy by 14%. The linear threshold policy has a performance drop of 12% compared to the solution obtained using SARSOP. This can be attributed to the linear hyperplane approximation to the threshold curve compared to the SARSOP solution where the number of linear segments is exponential in the number of states and observations.

Although the linear threshold policies have a slight performance drop compared to SARSOP, it has two significant advantages: (i) The policy (the linear threshold vectors corresponding to each stop) is easy to implement. In comparison, the SARSOP policy has approximately 7e4 piecewise linear segments. (ii) Computing the linear threshold approximation is computationally cheaper compared to SARSOP algorithm. It can be noted from Table 3 that Algorithm 1 is computationally cheaper by a factor of 10. The linear threshold policies that exploit the structure of the optimal policy perform nearly as well as the optimal policy computed via a general purpose approximate POMDP solver, with substantially lower computational cost.

Table 3
Comparison of the expected cumulative reward (Normalized w.r.t SARSOP) and number of computations by various algorithm. The linear threshold policies have a performance drop of 12% compared to the solution obtained using SARSOP and outperforms the heuristic policy by 14%. SARSOP solution computed using a 2.5 GHz CPU running for 2 hours. The calculation assumes a floating point operation every CPU cycle. Algorithm 1, for obtaining linear threshold policies, was run with finite horizon $N = 1000$.

Algorithm	Cumulative reward	#Computations
SARSOP	1	18e12
Linear Threshold	0.88	1.25e11
Heuristic	0.74	1.25e11
Periodic	0.35	0

6. Conclusion

We presented four main results regarding the multiple stopping time problem.

(i) The optimal policy was shown to be monotone with respect to a specialized monotone likelihood ratio order on lines (under reasonable conditions). Therefore the optimal policy was characterized by multiple threshold curves on the belief space and the optimal stopping sets satisfied a nested property (Theorem 1).

(ii) The cumulative reward was shown to be monotone with respect to the copositive ordering of the transition matrix (Theorem 2).

(iii) Necessary and sufficient conditions were given for linear threshold policies to satisfy the MLR increasing condition for the optimal policy (Theorems 3 and 4). We then gave a stochastic gradient algorithm (Algorithm 1) to estimate the linear threshold policies.

(iv) Finally, the linear scheduling policy was illustrated on a real dataset involving interactive advertising in live social media videos.

Extension of the current work involves developing upper and lower myopic bounds to the optimal policy as in Krishnamurthy and Pareek (2015), optimizing the ad length, and constraints on ad placement in the advertisement scheduling problem, multiple stopping problems with social learning (Krishnamurthy, 2012), and multiple stopping problems with measurement cost (Krishnamurthy, 2013).

Appendix A. Preliminaries and definitions

The proof of the main results require concepts in stochastic dominance (Karlin & Rinott, 1980) and submodularity (Topkis, 2011).

A.1. First-order and MLR stochastic dominance

To compare belief states, we use the monotone likelihood ratio (MLR) stochastic ordering and a specialized version of the MLR order restricted to lines in the simplex.

Definition 1 (*MLR Ordering*)**.** Let $\pi_1, \pi_2 \in \Pi$ be two belief state vectors. Then, π_1 is greater than π_2 with respect to Monotone Likelihood Ratio (MLR) ordering–denoted as $\pi_1 \geq_r \pi_2$, if

$$\pi_1(j)\pi_2(i) \leq \pi_2(j)\pi_1(i), \;\; i < j, \;\; i, j \in \{1, \ldots, S\} \tag{A.1}$$

Definition 2 (*First Order Stochastic Dominance*)**.** Let $\pi_1, \pi_2 \in \Pi$ be two belief state vectors. Then, π_1 is greater than π_2 with respect to first-order stochastic dominance–denoted as $\pi_1 \geq_s \pi_2$, if $\sum_{i=j}^{S} \pi_1(i) \leq \sum_{i=j}^{S} \pi_2(i), j \in \{1, 2, \ldots, S\}$.

Result (Krishnamurthy, 2016)*:*

(i) $\pi_1, \pi_2 \in \Pi$. Then, $\pi_1 \geq_r \pi_2$ implies $\pi_1 \geq_s \pi_2$.

(ii) $\pi_1 \geq_s \pi_2$ if and only if for any increasing function $\phi(\cdot)$, $\mathbb{E}_{\pi_1}\{\phi(x)\} \geq \mathbb{E}_{\pi_2}\{\phi(x)\}$.

For state-space dimension $S = 2$, MLR is a complete order and coincides with first-order stochastic dominance. For state-space dimension $S > 2$ MLR is a *partial order* i.e. $[\Pi, \geq_r]$ is a partially ordered set since it is not always possible to order any two belief states. However, on line segments in the simplex defined below, MLR is a total ordering.

Define the sub simplex $\mathcal{H}_i = \{\bar{\pi} : \bar{\pi} \in \Pi \text{ and } \bar{\pi}(i) = 0\}$, $i = 1, S$. Fig. 2 illustrates $\mathcal{H}_1$ for $S = 3$. Consider two types of lines, $\mathcal{L}(e_i, \bar{\pi})$; $i = 1, S$, where e_i is the unit indicator vector with 1 in the i position and 0 elsewhere, as follows: For any $\bar{\pi} \in \mathcal{H}_i$, construct the line $\mathcal{L}(e_i, \bar{\pi})$ that connects $\bar{\pi}$ to e_i as $\mathcal{L}(e_i, \bar{\pi}) = \{\pi : \pi = (1-\gamma)\bar{\pi} + \gamma e_i, \gamma \in [0, 1]\}$. For brevity, we denote $\mathcal{L}(e_i, \bar{\pi})$ by $\mathcal{L}(e_i)$. Fig. 2 illustrates the definition of $\mathcal{L}(e_1)$.

Definition 3 (*MLR Ordering on Lines*). π_1 is greater than π_2 with respect to MLR ordering on the lines $\mathcal{L}(e_i)$, denoted as $\pi_1 \geq_{\mathcal{L}_i} \pi_2$, if $\pi_1, \pi_2 \in \mathcal{L}(e_i)$, for some $\bar{\pi} \in \mathcal{H}_i$ and $\pi_1 \geq_r \pi_2$.

Remark 3 (*Krishnamurthy, 2016*). For $i = 1, S$, $\pi_1 \geq_{\mathcal{L}_i} \pi_2$ is equivalent to $\pi_j = \varepsilon_j e_i + (1-\varepsilon_j)\bar{\pi}$, for some $\bar{\pi} \in \mathcal{H}_i$ and $\varepsilon_1 \geq \varepsilon_2$.

Discussion: The MLR ordering on lines is a complete order, i.e. it forms a *chain*, meaning that all elements $\pi_1, \pi_2 \in \mathcal{L}(e_i)$ are comparable, i.e. either $\pi_1 \geq_{\mathcal{L}_i} \pi_2$ or $\pi_2 \geq_{\mathcal{L}_i} \pi_1$; see Krishnamurthy (2016). The complete order on $\mathcal{L}(e_i, \bar{\pi})$; $i = 1, S$ allows us to give a threshold characterization of the optimal policy on the belief space.

Definition 4 (*TP2*). A stochastic matrix, A is Totally Positive of order 2 (TP2), if all the second order minors are non-negative i.e. the determinants

$$\begin{vmatrix} a_{i_1 j_1} & a_{i_1 j_2} \\ a_{i_2 j_1} & a_{i_2 j_2} \end{vmatrix} \geq 0, \forall i_2 \geq i_1, j_2 \geq j_1 \tag{A.2}$$

An important consequence of (A1) and (A2) is the following theorem, which state that the Bayesian update $T(\pi, y)$ in (7) preserves MLR dominance.

Theorem 5 (*Krishnamurthy, 2016*). *If the transition matrix, P, and the observation matrix, B, satisfies the condition in* (A1) and (A2), *then*

- *For $\pi_1 \geq_r \pi_2$, the filter satisfies $T(\pi_1, \cdot) \geq_r T(\pi_2, \cdot)$.*
- *For $\pi_1 \geq_r \pi_2$, $\sigma(\pi_1, \cdot) \geq_s \sigma(\pi_2, \cdot)$*

Definition 5 (*Submodular Function*). A function $f : \mathcal{L}(e_i) \times \{1, 2\} \to \mathbb{R}$ is submodular if $f(\pi, u) - f(\pi, \bar{u}) \leq f(\bar{\pi}, u) - f(\bar{\pi}, \bar{u})$ for $u \geq \bar{u}$, $\pi \geq_{\mathcal{L}_i} \bar{\pi}$.

Theorem 6 (*Topkis, 2011*). *If $f(\pi, u)$ is submodular, then there exists a version of the optimal policy $u^*(\pi) = \arg\max_{u \in \mathcal{U}} f(\pi, u)$ that is decreasing in π.*

Hence, to prove the structural result, we show that the $Q(\pi, l, u)$ in (9) is submodular on the lines $\mathcal{L}(e_i)$; $i = 1, S$ with respect to the MLR order $\geq_{\mathcal{L}_i}$.

Appendix B. Value iteration

The value iteration algorithm is a successive approximation approach for solving Bellman's equation (9). However, in this paper, we use the value iteration algorithm in a mathematical induction proof; and not as a numerical algorithm. For iterations $k = 0, 1, \ldots,$

$$V_{k+1}(\pi, l) = \max_{u \in \{1,2\}} Q_{k+1}(\pi, l, u), \tag{B.1}$$

$$\mu_{k+1}(\pi, l) = \arg\max_{u \in \{1,2\}} Q_{k+1}(\pi, l, u), \tag{B.2}$$

where

$$Q_{k+1}(\pi, l, 1) = r'\pi + \rho \sum_y V_k(T(\pi, y), l-1)\sigma(\pi, y), \tag{B.3}$$

$$Q_{k+1}(\pi, l, 2) = \rho \sum_y V_k(T(\pi, y), l)\sigma(\pi, y), \tag{B.4}$$

with $V_0(\pi, l)$ initialized arbitrarily. Define $W_k(\pi, l)$ as

$$W_k(\pi, l) \triangleq V_k(\pi, l) - V_k(\pi, l-1). \tag{B.5}$$

The stopping and continue sets (at each iteration k) when l stops are remaining is defined as follows:

$$\begin{aligned} S^l_{k+1} &= \{\pi \,|\, r'\pi \geq \rho \sum_y W_k(T(\pi, y), l)\sigma(\pi, y)\}, \\ C^l_{k+1} &= \{\pi \,|\, r'\pi < \rho \sum_y W_k(T(\pi, y), l)\sigma(\pi, y)\}. \end{aligned} \tag{B.6}$$

The optimal stationary policy $\mu^*(\pi, l)$ is given by $\mu^*(\pi, l) = \lim_{k\to\infty} \mu_k(\pi, l)$.. Correspondingly, the stationary stopping and continue sets in (10) and (11) are given by

$$S^l = \lim_{k\to\infty} S^l_k, \quad C^l = \lim_{k\to\infty} C^l_k. \tag{B.7}$$

The value function, $V_k(\pi, l)$ in (B.1), can be rewritten, using (B.6), as follows:

$$\begin{aligned} V_k(\pi, l) = &\left(r'\pi + \rho \sum_y V_{k-1}(T(\pi, y), l-1)\sigma(\pi, y) \right) \mathcal{I}_{S^l_k} \\ &+ \left(\rho \sum_y V_{k-1}(T(\pi, y), l)\sigma(\pi, y) \right) \mathcal{I}_{C^l_k}, \end{aligned} \tag{B.8}$$

where $\mathcal{I}_{C^l_k}$ and $\mathcal{I}_{S^l_k}$ are indicator functions on the continue and stopping sets respectively, for each iteration k.

Assume $S^{l-1}_k \subset S^l_k$ (see Theorem 9) and substituting (B.8) in the definition of $W_k(\pi, l)$ in (B.5),

$$\begin{aligned} W_k(\pi, l) = &\left(\rho \sum_y W_{k-1}(T(\pi, y), l)\sigma(\pi, y) \right) \mathcal{I}_{C^l_k}(\pi) \\ &+ r'\pi \mathcal{I}_{C^{l-1}_k \cap S^l_k}(\pi) \\ &+ \left(\rho \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y) \right) \mathcal{I}_{S^{l-1}_k}(\pi). \end{aligned} \tag{B.9}$$

In order to prove the main theorem (Theorem 1), we require the following results, proofs of which are provided in Appendix C.

Theorem 7. *$V_k(\pi, l)$ is increasing in π.*

Theorem 8. *$W_k(\pi, l)$ is decreasing in l.*

Theorem 9. *$S^l_{k+1} \supset S^{l-1}_{k+1}$*

Appendix C. Proof of theorems

C.1. Proof of Theorem 7

Recall from (B.1), $V_k(\pi, l) = \max_{u \in \{1,2\}} Q_k(\pi, l, u)$. To prove Theorem 7, we show $Q_k(\pi, l, u)$ is MLR increasing in π for $u = \{1, 2\}$. From (B.3),

$$Q_k(\pi, l, 1) = r'\pi + \rho \sum_y V_{k-1}(T(\pi, y), l-1)\sigma(\pi, y),$$

Using Theorem 5 and the induction hypothesis, the term $\sum_y V_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)$ is MLR increasing in π. From (A3),

$r'\pi$ is MLR increasing in π. The proof for $Q_k(\pi, l, 2)$ MLR increasing in π is similar and is omitted. Hence, $V_k(\pi, l)$ is MLR increasing in π.

C.2. Proof of Theorem 8

The proof follows by induction. From (B.9), we have

$$W_k(\pi, l-1) = \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)\mathcal{I}_{C_k^{l-1}}(\pi) + r'\pi\mathcal{I}_{C_k^{l-2}\cap S_k^{l-1}}(\pi) + \sum_y W_{k-1}(T(\pi, y), l-2)\sigma(\pi, y)\mathcal{I}_{S_k^{l-2}}(\pi) \quad \text{(C.1)}$$

Hence, we compare $W_k(\pi, l)$ and $W_k(\pi, l-1)$ in the following 4 regions:

(a) S_k^{l-2} : $W_k(\pi, l) - W_k(\pi, l-1) = \sum_y (W_{k-1}(T(\pi, y), l-1) - W_{k-1}(T(\pi, y), l-2))\sigma(\pi, y)$, which is non-negative by the induction assumption.
(b) $C_k^{l-2}\cap S_k^{l-1}$: $W_k(\pi, l) - W_k(\pi, l-1) = \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y) - r'\pi$, which is non-negative since $\pi \in S_k^{l-1}$.
(c) $C_k^{l-1}\cap S_k^{l}$: $W_k(\pi, l) - W_k(\pi, l-1) = r'\pi - \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)$, which is non-negative since $\pi \in C_k^{l-1}$.
(d) C_k^{l}: Similar to Case a above.

C.3. Proof of Theorem 9

If $\pi \in S_k^{l-1}$, then $r'\pi \geq \sum_y W_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)$. By Theorem 8, $r'\pi \geq \sum_y W_{k-1}(T(\pi, y), l)\sigma(\pi, y)$. Hence $\pi \in S_k^{l}$.

C.4. Proof of Theorem 1

Existence of optimal policy: In order to show the existence of a threshold policy of $\mathcal{L}(e_1)$, we need to show that $Q_{k+1}(\pi, l, 2) - Q_{k+1}(\pi, l, 1)$ is submodular in $\pi \in \mathcal{L}(e_1)$. Since, $Q_{k+1}(\pi, l, 2) - Q_{k+1}(\pi, l, 1) = \rho\sum_y W_k(T(\pi, y), l)\sigma(\pi, y) - r'\pi$. We need to show that $\rho\sum_y W_k(T(\pi, y), l)\sigma(\pi, y) - r'\pi$ is MLR decreasing in π.

$$\begin{aligned}
&\rho\sum_y W_k(T(\pi, y), l)\sigma(\pi, y) - r'\pi \quad \text{(C.2)}\\
&= \sum_y \left(\left(\rho W_k(T(\pi, y), l) - \rho r'T(\pi, y)\right)\right.\\
&\left. - \left(r'\pi - \rho r'T(\pi, y)\right)\right)\sigma(\pi, y)\\
&= \rho\sum_y \left(W_k(T(\pi, y), l) - r'T(\pi, y)\right)\sigma(\pi, y)\\
&- r'(I - \rho P')\pi \quad \text{(C.3)}
\end{aligned}$$

The term $-r'(I - \rho P')\pi$ in (C.3) is MLR decreasing in π due to our assumption. Hence, to show that $\rho\sum_y W_k(T(\pi, y), l)\sigma(\pi, y) - r'\pi$ is MLR decreasing in π it is sufficient to show that $W_k(\pi, l) - r'\pi$ is MLR decreasing in π. Define, $\bar{W}_k(\pi, l) \triangleq W_k(\pi, l) - r'\pi$.

Now, $\bar{W}_k(\pi, l) =$

$$\left(\sum_y \rho\left(\left(\bar{W}_{k-1}(T(\pi, y), l) + r'T(\pi, y)\right) - r'\pi\right)\sigma(\pi, y)\right)\mathcal{I}_{C_k^l}(\pi) +$$

$$\left(\sum_y \rho\left(\left(\bar{W}_{k-1}(T(\pi, y), l-1) + r'T(\pi, y)\right) - r'\pi\right)\sigma(\pi, y)\right)\mathcal{I}_{S_k^l}(\pi)$$

$$= \left(\sum_y \left(\rho\bar{W}_{k-1}(T(\pi, y), l)\sigma(\pi, y)\right) - r'(I - \rho P)'\pi\right)\mathcal{I}_{C_k^l}(\pi) +$$

$$\left(\sum_y \left(\rho\bar{W}_{k-1}(T(\pi, y), l-1)\sigma(\pi, y)\right) - r'(I - \rho P)'\pi\right)\mathcal{I}_{S_k^l}(\pi) \quad \text{(C.4)}$$

We prove using induction that $\bar{W}_k(\pi, l)$ is MLR decreasing in π, using the recursive relation over k in (C.4). For $k = 0$, $\bar{W}_0(\pi, l) = W_0(\pi, l) - r'\pi = V_0(\pi, l) - V_0(\pi, l-1) - r'\pi$. The initial conditions of the value iteration algorithm can be chosen such that $\bar{W}_0(\pi, l)$ is decreasing in π.

Next, we show that $\bar{W}_k(\pi, l)$ is MLR decreasing in π, if $\bar{W}_{k-1}(\pi, l)$ is MLR decreasing in π. For $\pi_1 \geq_r \pi_2$, consider the following cases: (a) $\pi_1, \pi_2 \in S_k^{l-1}$, (b) $\pi_1 \in S_k^{l-1}$, $\pi_2 \in C_k^{l-1}\cap S_k^l$, (c) $\pi_1, \pi_2 \in C_k^{l-1}\cap S_k^l$, (d) $\pi_1 \in C_k^{l-1}\cap S_k^l$, $\pi_2 \in C_k^l$, (e) $\pi_1, \pi_2 \in C_k^l$, (f) $\pi_1 \in S_k^{l-1}$, $\pi_2 \in C_k^l$. For cases (a), (c), (e), $\bar{W}_k(\pi_1, l) \leq \bar{W}_k(\pi_2, l)$ by the induction assumption. For case (b) $\bar{W}_k(\pi_1, l) \leq \bar{W}_k(\pi_2, l)$, since $\pi_1 \in S_k^{l-1}$. Case (d) is similar to case (b). For case (f), $\bar{W}_k(\pi_1, l) - \bar{W}_k(\pi_2, l) =$

$$\begin{aligned}
&\left(\sum_y \left(\rho\bar{W}_{k-1}(T(\pi_1, y), l-1)\sigma(\pi_1, y)\right) - r'(I - \rho P)'\pi_1\right)\\
&- \left(\sum_y \left(\rho\bar{W}_{k-1}(T(\pi_2, y), l)\sigma(\pi_2, y)\right) - r'(I - \rho P)'\pi_2\right)\\
&\leq \rho\left(\sum_y \left(\left(\bar{W}_{k-1}(T(\pi_1, y), l-1) - \bar{W}_{k-1}(T(\pi_1, y), l)\right)\sigma(\pi_1, y)\right)\right)\\
&\leq 0,
\end{aligned}$$

where the first inequality is due to induction hypothesis and the second inequality is due to Theorem 8. Hence, $\bar{W}_k(\pi, l)$ is decreasing in π, if $\bar{W}_{k-1}(\pi, l)$ is decreasing in π, finishing the induction step.

Characterization of the switching curve Γ_l: For each $\bar{\pi} \in \mathcal{H}$ construct the line segment $\mathcal{L}(e_1, \bar{\pi})$. The line segment can be described as $(1-\varepsilon)\bar{\pi} + \varepsilon e_1$. On the line segment $\mathcal{L}(e_1, \bar{\pi})$ all the belief states are MLR orderable. Since $\mu^*(\pi, l)$ is monotone decreasing in π, for each l, we pick the largest ε such that $\mu^*(\pi, l) = 1$. The belief state, $\pi^{\varepsilon^*, \bar{\pi}}$ is the threshold belief state, where $\varepsilon^* = \inf\{\varepsilon \in [0, 1] : \mu^*(\pi^{\varepsilon, \bar{\pi}}, l) = 1\}$. Denote by $\Gamma_l(\bar{\pi}) = \pi^{\varepsilon^*, \bar{\pi}}$. The above construction implies that there is a unique threshold $\Gamma_l(\bar{\pi})$ on $\mathcal{L}(e_1, \bar{\pi})$. The entire simplex can be covered by considering all pairs of lines $\mathcal{L}(e_1, \bar{\pi})$, for $\bar{\pi} \in \mathcal{H}_1$, i.e. $\Pi = \cup_{\bar{\pi}\in\mathcal{H}}\mathcal{L}(e_1, \bar{\pi})$. Combining, all points yield a unique threshold curve in Π given by $\Gamma_l = \cup_{\bar{\pi}\in\mathcal{H}_1}\Gamma_l(\bar{\pi})$.

Connectedness of S^l and C^l: Since $e_1 \in S^l$ for all l, call S_a^l, the subset of S^l that contains e_1. Suppose S_b^l is the subset that was disconnected from S_a^l. Since every point on Π lies on the line segment $\mathcal{L}(e_1, \bar{\pi})$, for some $\bar{\pi}$, there exists a line segment starting from $e_1 \in S_a^l$ that would leave the set S_a^l, pass through the set where action 2 is optimal and then intersect set S_b^l, where action 1 is optimal. But, this violates the requirement that the policy $\mu^*(\pi, l)$ is monotone on $\mathcal{L}(e_1, \bar{\pi})$. Hence, S_a^l and S_b^l are connected. The proof for connectedness of C^l is similar to S^l by considering the line $\mathcal{L}(e_S, \bar{\pi})$ instead of $\mathcal{L}(e_1, \bar{\pi})$, and is hence omitted.

Nested structure: The proof follows from Theorem 9.

C.5. Copositive ordering

Definition 6 (*Copositive Ordering (Krishnamurthy, 2016)*)**.** Given two $S \times S$ transition matrices P and Q, we say that $P \preceq Q$ if the sequence of $S \times S$ matrices Γ^j are: $\pi'\Gamma^j\pi \geq 0, \pi \in \Pi, j = 1, \ldots, S-1$, where each element of Γ^j is given by $\Gamma^j_{m,n} = \frac{1}{2}\left(\gamma^j_{m,n} + \gamma^j_{n,m}\right)$, $\gamma^j_{m,n} = P_{m,j}Q_{n,j+1} - P_{m,j+1}Q_{n,j}$.

Theorem 10 (*Krishnamurthy, 2016, Theorem 10.6.1*). *Suppose transition matrices $\underline{P}$ and $\bar{P}$ are constructed such that $\underline{P} \preceq P \preceq \bar{P}$. Then for any observation y and belief $\pi \in \Pi$, the filtering update $T(\pi, y; P)$*[11] *in (7) satisfies $T(\pi, y; \underline{P}) \leq_r T(\pi, y; P) \leq_r T(\pi, y; \bar{P})$.*

C.6. Proof of Theorem 4

For $l_1 > l_2$, due to the nested structure in Theorem 1 $S^{l_2} \subset S^{l_1}$. This implies $\mu_\theta(l_2, \pi) \geq \mu_\theta(l_1, \pi)$, i.e., $\begin{bmatrix} 0 & 0 & \theta_{l_2} - \theta_{l_1} \end{bmatrix} \begin{bmatrix} \pi \\ -1 \end{bmatrix} \geq 0$. It is straightforward to check that the conditions in (16) in Theorem 4 satisfy the above conditions.

Appendix D. Proof of propositions

D.1. Proof of Proposition 1

Let $v = (I - \rho P)r$. When $\rho < 1$, $(I - \rho P)$ is invertible. Hence, $r = (I - \rho P)^{-1} v = \sum_{k=0}^{\infty} \rho^k P^k v$. Since the product of TP2 matrices is TP2, each P^k is TP2. Then, r been decreasing follows from Theorem 9.2.2 in Krishnamurthy (2016). For $\rho = 1$, $g = \lim_{\rho \uparrow 1} (1 - \rho)(I - \rho P)^{-1} v$ is the solution of $(I - P)r = v$. This limit exists (Puterman, 2005, Cor. 8.2.5) and hence, r has decreasing elements.

D.2. Proof of Proposition 2

The proof follows from the finite stopping time property of the multiple stopping time problem; see Footnote 2. A finite horizon POMDP with a finite state and observation space has a value function that is piecewise linear and convex; see Theorem 7.4.1 in Krishnamurthy (2016). For $l = 1$, $V(\pi, 1) = \max_{\gamma \in \Gamma} \gamma' \pi$, where Γ is a finite set due to the finite stopping time property. For $l = 2$, the dynamic programming equation in (9) can be written as:

$$V(\pi, 2) = \max \left\{ r'\pi + \max_{\gamma \in \Gamma} \gamma' P' \pi, \rho \sum_{y \in \mathcal{Y}} V(T(\pi, y), 2)\, \sigma(\pi, y) \right\}.$$

For each $\gamma \in \Gamma$, the stopping set is convex; see the proof of Theorem 12.2.1 in Krishnamurthy (2016). Hence, the stopping set for $l = 2$ is a union of convex sets. Similar argument holds for any value of l.

Appendix E. Finite horizon algorithms

Algorithm 2 details the steps to compute the finite time horizon approximation in (17) for the linear threshold policies. Its input is the POMDP parameters, policy (in terms of the parameter θ) and number of stops. It computes the accumulated reward using the input policy by running a POMDP simulation of at most N time points. Algorithm 3 summarizes the computation of the finite time horizon approximation with the softmax parametrization of the policy in (23). The key difference with Algorithm 2 is in Steps 5–7: the softmax policy in (23) replaces the linear threshold policies in Step 5 of Algorithm 2.

[11] The notation $T(\cdot, \cdot; P)$ makes explicit the transition matrix used in the filter update.

Algorithm 2 Finite Horizon Approximation Algorithm for linear threshold policies

Require: Finite time approximation parameter N, policy parameter θ, number of stops L, initial belief π_0, discount factor ρ, reward vector r.
1: $l \leftarrow L, J \leftarrow 0$.
2: **for** iterations $n = 1, 2, \cdots N$: **do**
3: **while** $l \neq 0$ **do**
4: Obtain observation Y_n and update belief π_n according to (7).
5: Compute $a_n \leftarrow \mu_\theta(\pi_n, l)$ according to (15).
6: **if** $a_n = 1$ **then**
7: $J \leftarrow J + \rho^n \pi_n' r, l \leftarrow l - 1$
return J

Algorithm 3 Finite Horizon Approximation Algorithm: Using softmax parametrization

Identical to Algorithm 2 except for Steps 5-7
5: actionprob $= \left[\exp\left(\begin{bmatrix} 0 & \theta_{l,1} \end{bmatrix}' \pi \right) \exp\left(\begin{bmatrix} 0 & \theta_{l,2} \end{bmatrix}' \pi \right) \right]$
6: actionprob $\leftarrow$ actionprob$/\sum$ actionprob
7: Sample $a_n \sim$ actionprob

References

Alexander, S. H. J., & Nikolaev, G. (2010). Stochastic sequential decision-making with a random number of jobs. *Operations Research*, *58*(4), 1023–1027.

Baldominos, A., Esperanza, A., Marrero, I., Saez, Y. (2016). Real-time prediction of gamers behavior using variable order Markov and big data technology: A case study.

Bayraktar, E., & Kravitz, R. (2015). Quickest detection with discretely controlled observations. *Sequential Analysis*, *34*(1), 77–133.

Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proc. of IMC* (pp. 49–62).

Carmona, R., & Touzi, N. (2008). Optimal multiple stopping and valuation of swing options. *Mathematical Finance*, *18*(2), 239–268.

Dahlgren, E., & Leung, T. (2015). An optimal multiple stopping approach to infrastructure investment decisions. *Journal of Economic Dynamics & Control*, *53*, 251–267.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Geng, J., Bayraktar, E., & Lai, L. (2014). Bayesian quickest change-point detection with sampling right constraints. *IEEE Transactions on Information Theory*, *60*(10), 6474–6490.

Hamilton, W.A., Garretson, O., Kerne, A. (2014). Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Proc. of the ACM conf. on human factors in computing systems* (pp. 1315–1324).

Kang, H., & McAllister, M. P. (2011). Selling you and your clicks: examining the audience commodification of google. *Journal for a Global Sustainable Information Society*, *9*(2), 141–153.

Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. *Journal of Multivariate Analysis*, *10*(4), 467–498.

Krasnosielska-Kobos, A. (2015). Multiple-stopping problems with random horizon. *Optimization*, *64*(7), 1625–1645.

Krishnamurthy, V. (2011). Bayesian sequential detection with phase-distributed change time and nonlinear penalty – A POMDP lattice programming approach. *IEEE Transactions on Information Theory*, *57*(10), 7096–7124.

Krishnamurthy, V. (2012). Quickest detection POMDPs with social learning: Interaction of local and global decision makers. *IEEE Transactions on Information Theory*, *58*(8), 5563–5587.

Krishnamurthy, V. (2013). How to schedule measurements of a noisy Markov chain in decision making?. *IEEE Transactions on Information Theory*, *59*(7), 4440–4461.

Krishnamurthy, V. (2016). *Partially observed Markov decision processes*. Cambridge University Press.

Krishnamurthy, V., & Bhatt, S. (2016). Sequential detection of market shocks with risk-averse CVaR social sensors. *IEEE Journal of Selected Topics in Signal Processing*, *10*(6), 1061–1072.

Krishnamurthy, V., & Pareek, U. (2015). Myopic bounds for optimal policy of POMDPs: An extension of Lovejoy's structural results. *Operations Research*, *62*(2), 428–434.

Krishnamurthy, V., & Rojas, C. R. (2014). Reduced complexity HMM filtering with stochastic dominance bounds: A convex optimization approach. *IEEE Transactions on Signal Processing*, *62*(23), 6309–6322.

Kurniawati, H., Hsu, D., & Lee, W. S. (2008). SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*.

Lai, T. L. (1997). On optimal stopping problems in sequential hypothesis testing. *Statistica Sinica*, *7*(1), 33–51.
Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement. In *Int. conf. on user modeling, adaptation, and personalization* (pp. 164–175). Springer.
Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, *109*(1), 68–72.
Lobel, I., Patel, J., Vulcano, G., & Zhang, J. (2015). Optimizing product launches in the presence of strategic consumers. *Management Science*, *62*(6), 1778–1799.
Müller, A., & Stoyan, D. (2002). *Comparison methods for stochastic models and risks*. Chichester: John Wiley&Sons Ltd.
Nakai, T. (1985). The problem of optimal stopping in a partially observable Markov chain. *Journal of Optimization Theory and Applications*, *45*(3), 425–442.
Nikolaev, M. (1999). On optimal multiple stopping of Markov sequences. *Theory of Probability & Its Applications*, *43*(2), 298–306.
Poor, H. V., & Hadjiliadis, O. (2008). *Quickest detection*. Cambridge University Press.
Popescu, D. G., & Crama, P. (2015). Ad revenue optimization in live broadcasting. *Management Science*, *62*(4), 1145–1164.
Puterman, M. L. (2005). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
Smith, T., Obrist, M., & Wright, P. (2013). Live-streaming changes the (video) game. In *Proc. of the 11th European conference on interactive TV and video* (pp. 131–138). ACM.
Spall, J.C. (2005). *Vol. 65: Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.
Stadje, W. (1987). An optimal k-stopping problem for the Poisson process. In *Mathematical statistics and probability theory* (pp. 231–244). Springer.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction. Vol. 1*. MIT Press Cambridge.
Topkis, D. M. (2011). *Supermodularity and complementarity*. Princeton university press.
Wang, B., Zhang, X., Wang, G., Zheng, H., & Zhao, B. Y. (2016). Anatomy of a personalized livestreaming system. In *Proc. IMC '16* (pp. 485–498). NY, USA: ACM.

Vikram Krishnamurthy is a professor at Cornell Tech and the School of Electrical and Computer Engineering, Cornell University. His current research interests are in partially observed Markov decision processes, social network design/analysis and stochastic optimization. He is author of the books "Partially Observed Markov Decision Processes" and "Dynamics of Artificial Membranes and Biosensors" published by Cambridge University Press in 2016 and 2018, respectively.

Anup Aprem obtained his M.E from Indian Institute of Science in 2012 and Ph.D. from the University of British Columbia in 2017. His research interests are statistical signal processing and decision making in the area of social networks and social media.

Sujay Bhatt is a Ph.D. student at Cornell University in the Electrical and Computer Engineering department. His research interests are in social learning and stochastic control. He obtained his M.Tech degree from Indian Institute of Technology Bombay (IIT-B), India.