Tracking of Real-Valued Markovian Random Processes with Asymmetric Cost and Observation

Parisa Mansourifard[1], Bhaskar Krishnamachari[1], and Tara Javidi[2]

***Abstract*— We study a state-tracking problem in which the background random process is Markovian with unknown real-valued states and known transition probability densities. At each time step the decision-maker chooses a state as an action and accumulates some reward based on the selected state and the actual state. If the selected state is higher than the actual state, the actual state is fully observed in expense of over-utilization cost. Otherwise, the decision-maker has to pay under-utilization cost and could only observe the actual state partially (that it is higher than the selected state). Thus, the decision-maker faces asymmetries in both cost and observation. The goal is to select the actions in order to maximize the total expected discounted reward over infinite horizon. We model this problem as a Partially Observable Markov Decision Process and formulate it in two different ways: (i) belief-based, and (ii) sequence-based. In the sequence-based formulation, only two parameters matter to define the sequence of actions, the last fully observed state and the time passed from the last observation. We prove key structural properties of the optimal policy including a lower bound on the optimal sequence. Further, for a specific form of processes we present an upper bound on the optimal sequence. Both lower and upper bound sequences have percentile threshold structure and are monotonically increasing with respect to the last fully observed state.**

I. Introduction

In many network protocols, the devices must set the communication parameters to maximize the utilization of the resource whose availability is a stochastic process. One prominent example is congestion control, in which a transmitter must select the transmission rate to utilize the available bandwidth, which varies randomly due to the dynamic nature of traffic load imposed by other users on the network [1], [2]. Another example is in a communication system where the transmitter must select the transmission rate in order to maximize the number of successfully transmitted bits [3].

Structure of optimal policies has been established for simpler related problems of optimizing transmissions over a two-state Gilbert-Elliott channel in [3], [4]. In this work, we consider a more general case of real-valued Markovian channel. Our recent related work [1] is about a Bayesian congestion control problem with a discrete-state space where a source must select a transmission rate at each time step over a network with a Markovian available bandwidth such that the less congestion occurs (less over-utilization cost) and more information about the actual bandwidth reveals. In this example, the bandwidth maps to the actual state of the background random process and the transmission rate maps to the selected state.

In this paper we consider a generalized version of the problem where the actual state of the background Markovian random process could be any real value in a defined range. We assume that the transition probability densities for the background process are known but the actual state is not fully observable. The goal it to select a state as an action at each time step in order to maximize the total expected discounted reward over infinite horizon. The reward accumulated at each time step is a piecewise linear function of the difference between the selected and the actual states. If the selected state is higher than the actual state, the decision-maker gets full observation about the actual state which is useful for future decision, but he has to pay an over-utilization penalty. The decision-maker may want to behave conservatively and select a lower state. But in this case he gets only partial observation about the state, that it is higher than the selected action. In this case, he has to pay under-utilization cost which is usually less than the over-utilization cost. Therefore, the decision-maker faces a trade-off between accumulating higher immediate reward and getting more information about the actual state.

We model this problem as a Partially Observable Markov Decision Process (POMDP) problem since the decision-maker does not have full observation about the actual state. This POMDP problem does not have an efficiently computable solution [5], *i.e.* the optimal policy which could provide the solution of the POMDP problem is not computationally tractable. We present key structural properties of the optimal policy as well as a new formulation of the problem based on the sequence of the actions. We show that the optimal policy can be perfectly characterized by only two parameters: (i) the last fully observed actual state (whenever the selected state is higher than the actual state, we get full observation), and (ii) the time steps passed since the last full observation. Therefore, instead of looking for the best action at each time step maximizing the expected reward-to-go, we can look for the best action sequence for each last fully observed state which will be followed up to the time step where the action is higher than the actual state. At this point the actual state is fully observed, namely the last fully observed state resets to a new value. After this point, we will continue with the optimal sequence corresponding to the new

*This work was supported in part by the U.S. National Science foundation under ECCS-EARS awards numbered 1247995 and 1248017, by the Okawa foundation through an award to support research on "Network Protocols that Learn", and a partial support from L3-communications as well as UCSD's center for Wireless Communications and Networked Systems.

[1]Parisa Mansourifard and Bhaskar Krishnamachari are with Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles 90089 CA USA. parisama,bkrishna@usc.edu

[2]Tara Javidi is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA. tjavidi@ucsd.edu

fully observed state. To the best of our knowledge, this work is the first to represent the sequence-based formulation.

We prove that each optimal sequence is lower bounded by the sequence of actions generated by the myopic policy starting from the same last fully observed state. The myopic policy at each time step selects an action which achieves the supremum of the immediate expected reward, ignoring its impact on the future reward. We also show that if transition probability densities preserve the First Order Stochastic Dominance (FOSD) on Probability Distribution Functions (PDF), the myopic policy is monotonically increasing with respect to the last fully observed state. In other words, for the higher last fully observed states the whole myopic sequence will be higher than the one starting from a lower last fully observed state. We show that the myopic policy has a percentile threshold structure for all transition probability densities. The percentile threshold structure means that the selected state is equal to the lowest state above a given percentile of the PDFs. Further, we consider a specific form of processes defined as Independent Increment Markov Chain (IIMC) (See Section VI for definition and [2] for more details). For these processes, we derive an upper bound on the optimal sequences with the assumption of zero under-utilization cost. We show that the upper-bound sequence also has a percentile threshold structure and follows the same monotonicity property.

II. Problem Formulation

We consider a discrete-time continuous-state Markovian process whose state is denoted by B_t. The transition probability densities are assumed to be known but the actual state of the background Markovian process is unknown. At each time step, the decision-maker selects a state, as an action, based on the history of observations and accumulates a reward as a piecewise linear function of the selected state and the actual state B_t.

The goal is to select the sequential actions which maximize the total expected discounted reward accumulated over the infinite horizon. We formulate our decision-making problem within a POMDP-based framework defined as follows:

- *State:* The actual state of the Markov process B_t at time step t, can be any real number in the range of $\mathcal{M} = [m, M]$, *i.e.* the state space.
- *State transition:* The transition probability densities of the actual states over time are shown $\forall m \leq x, y \leq M$ by

$$p(x|y) := P(B_t = x | B_{t-1} = y).$$

- *Action:* At each time step, we choose an action r_t from the action space which is equivalent to the state space $\mathcal{M}$.
- *Observed information:* The observed information at time step t is defined by the event $o_t(r_t) \in \mathcal{O}$ which will be useful for the decision at the next time step. The possible observations corresponding to the action r_t is as follows:

  - $o_t(r_t) = \{B_t = i\}, \forall i \in [m, r_t)$ is the event of fully observing the actual state B_t. This corresponds to the selection of the state higher than B_t.
  - $o_t(r_t) = \{B_t \geq r_t\}$ is the event of partial observing that B_t is larger than or equal to the selected state.
- *Reward:* The immediate reward earned at time step t is defined as follows:

$$R(B_t, r_t) = \begin{cases} qB_t - C_u(r_t - B_t) & \text{if } r_t > B_t \\ qr_t - C_l(B_t - r_t) & \text{if } r_t \leq B_t, \end{cases} \quad (1)$$

  where C_u and C_l are the over-utilization and the under-utilization cost coefficients, respectively, and q is the gain unit.

III. Related Work

We review some recent works in the literature dealing with similar problmes. Johnston and Krishnamurthy [4] consider the problem of minimizing the transmission energy and latency associated with transmitting a file across a Gilbert Elliott fading channel, formulate it as a POMDP, identify a threshold policy for it, and analyze it for various parameter settings. Laourine and Tong [3], consider betting on Gilbert Elliott Channels with three possible choices of actions, and shows that a threshold-type policy consisting of one, two, or three thresholds depending on the parameters, is optimal. Wu and Krishnamachari [6] study the optimal transmission policy for a Gillbert-Elliot channel with unknown statistics.

This problem is also known as Newsvendor problem with partially observed perishable inventory levels, in the context of operation management research. The newsvendor problem maps the demand to the background random process and the inventory level (how many items to store in order to satisfy the demand) to the action [7]. Most of the works done in the inventory management literature, *e.g.* [8], [9], assume that the demand process is independent and identically distributed (i.i.d) at different time steps. With this assumption, the optimal policy is exactly equal to the myopic policy. But here we assume that the background process is Markovian; thus the myopic policy provides only a lower bound on the optimal policy.

Bensoussan *et al.* [10] consider a Newsvendor problem with the assumption of Markovian demand process. They use the un-normalized beliefs to prove the existence of the optimal policy and show that the myopic policy provides a lower bound on the actions selected by optimal policy. In this paper, in contrast to their work, we introduced a sequence-based formulation and show that the optimal sequence is lower bounded by the sequence generated by the myopic policy. Further, by investigating a specific form of the transition probability densities, called IIMC, we derive an upper bound for the optimal sequence which also has a percentile threshold structure similar to the myopic sequence.

IV. Two Equivalent Value Iterations

We can represent our decision-making problem in two different ways. (i) Belief-based: We define our Prior Belief Distribution (PBD) as the probability density function (PDF)

of our beliefs about the states, shown by $f_t(x)$, at each time step and try to maximize the expected discounted reward-to-go corresponding to the PDF. (ii) Sequence-based: We formulate the problem based on the action sequences starting from each possible fully observed state and try to find the best sequence to maximize the total expected discounted reward. We consider both formulations and show that they are equivalent.

A. Belief-Based Value Iteration

In the belief-based formulation, the decision-maker keeps a belief about the probability distribution of the state space given all past observations, denoted by $f_t(x)$ where $x \in \mathcal{M}$ indicates the actual state, and selects the action based on the PBD. It can be shown that the belief is a sufficient statistic of the complete observation history (see *e.g.*, [11]).

The PBD updating for the next time step, upon the selected action r_t and the observation, is given $\forall x \in \mathcal{M}$ by:

$$f_{t+1}(x) = \begin{cases} \int_m^M T_{r_t}[f_t](\alpha)p(x|\alpha)d\alpha & \text{if } r_t \leq x_t \\ p(x|x_t) & \text{if } r_t \geq x_t \end{cases}, \quad (2)$$

where T_r is a non-linear operation on a PBD f, as follows:

$$T_r[f](x) = \begin{cases} 0 & \text{if } x < r \\ \frac{f(x)}{\int_r^M f(\alpha)d\alpha} & \text{if } x \geq r \end{cases}. \quad (3)$$

The immediate expected reward, achieved by selecting the action r_t and based on the PBD f_t is obtained by taking expectation of (1), as follows:

$$\begin{aligned} \bar{R}(f_t; r_t) &= \int_{x=m}^{M} f_t(x)R(x, r_t)dx \\ &= \int_{x=r_t}^{M} f_t(x)[qr_t - C_l(x - r_t)]dx \\ &+ \int_{i=m}^{r_t} f_t(x)[qx - C_u(r_t - x)]dx. \end{aligned} \quad (4)$$

The goal is to maximize the total expected discounted reward over all admissible policies π, given by

$$\max_{\pi} J^{\pi}(f_0) = \max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R(B_t; r_t)|f_0], \quad (5)$$

where $0 \leq \beta < 1$ denotes the discount factor and f_0 is the initial PBD. $J^{\pi}(f_0)$ is the total expected discounted reward accumulated over the infinite horizon under policy π and starting in the initial PBD f_0. The policy π specifies a sequence of functions $\pi_1, \pi_2, ...$, where π_t maps a PBD f_t to an action at time step t, *i.e.*, $r_t = \pi_t(f_t)$. The optimal policy denoted by π^{opt} is a policy which maximizes (5). This problem may be solved using the following fixed point equations:

$$V(f_t) = \sup_{r_t} V(f_t; r_t), \quad (6)$$

$$\begin{aligned} V(f_t; r_t) &= \bar{R}(f_t; r_t) + \beta \int_m^{r_t} V(p(x|x_t))f_t(x_t)dx_t \\ &+ \beta V(\int_{r_t}^{M} T_{r_t}[f_t](\alpha)p(x|\alpha)d\alpha) \int_{r_t}^{M} f_t(x_t)dx_t, \end{aligned}$$

where sup is the notation for the supremum. The existence of the optimal policy for the above value iteration is proved in [10]. A policy π^{opt} is optimal if for $t = 1, 2, ...$; $r_t^{opt}(f_t)$ achieves the maximum in (6), denoted by:

$$r_t^{opt}(f_t) := \arg\max_{r \in \mathcal{M}} V(f_t; r). \quad (7)$$

B. Sequence-Based Value Iteration

In the sequence-based formulation, instead of the action for each PBD, the decision-maker makes his decision about the whole action sequence starting from any fully observed state. We can formulate the problem in this way because the optimal policy can also be perfectly characterized by only two parameters; (i) the last fully observed state, namely s_L, and (ii) the time steps passed since the last observation, say t_L. In other words, for each s_L there exists an optimal sequence which can be followed up to the nest full observation where the action is higher than the actual state and s_L will be reset to the new full observed state.

Now let us denote the sequence of actions starting from state i by $a(i, .) = \{a(i, 1), a(i, 2), ...\}$ where $a(i, t_L)$ is the action selected at t_L time steps passed from the last fully observed state i. An example of action sequences taken by an arbitrary policy and a sample path of the Markovian random process is shown in Fig. 1. Let the policy follows the shifted version of the action sequence $a(0, .)$ after any full observation, *i.e.* if the state i is fully observed, the policy will follow the action sequence of $a(0, .) + i$. Let assume the actual state at $t = 0$ is fully observed. Therefore, the action sequence corresponding to the initial point ($s_L = 2$) which is $a(0, .) + 2$ is followed up to a point where the action sequence exceeds the sample path. At this point ($t = 6$) the actual state is fully observed and the sequence will be reset to the actual state ($s_L = 3.25$). After this point, the sequence corresponding to the new fully observed state $a(3.25, .) = a(0, .) + 3.25$ is followed. At the reset point the over-utilization cost occurs and at the other time steps the under-utilization costs have to be paid.

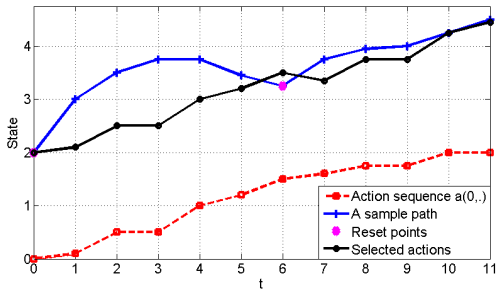

Fig. 1. An example of executing an arbitrary policy on a sample path of the Markovian process.

The goal of the decision-maker is to find the best policy of sequences in order to maximize the total expected discounted reward. The supremum of the total expected discounted reward collected from the last fully observed state $s_L = i$ is

given by:

$$W(i) = \sup_{a(i,t_L)\in[m,M],\forall t_L} W(i; a(i,.)), \tag{8}$$

$$W(i; a(i,.)) = \sum_{t_L=1}^{\infty} \int_{j=m}^{a(i,t_L)} P^{t_L}_{i,a(i,1:t_L-1),j} dj \times [\sum_{\tau=1}^{t_L-1} \beta^{\tau-1}((q+C_l)a(i,\tau) - C_l \bar{B}(i,\tau)) + \beta^{t_L-1}((q+C_u)j - C_u a(i,t_L)) + \beta^{t_L} W(j)], \tag{9}$$

where the term inside [.] given in (9) is the expected discounted reward accumulated conditioned on the occurrence of the following event: no reset (*i.e.* full observation) at time steps $1, 2, ..., t_L - 1$ passed from the last fully observed state $s_L = i$ and following the action sequence of $a(i, 1 : t_L) = \{a(i, 1), ..., a(i, t_L)\}$ and reset to the actual state j at t_L. The probability of occurrence of this event denoted by $P^{t_L}_{i,a(i,1:t_L-1),j}$ is given by (10) for $m \leq i, j \leq M$ and is 0 otherwise.

Note that $\bar{B}(i, \tau)$ is the mean of the actual state at time step τ passed from the last fully observed state i without any reset before τ, given by:

$$\bar{B}(i,\tau) = \int_{x=a(i,\tau)}^{M} x P^{\tau}_{i,a(i,1:\tau-1),x} dx.$$

$W(j)$ can also be computed recursively by substituting i with j in (8). The optimal sequence achieved by the above value iteration is given by:

$$a^{opt}(i,.) = \arg \sup_{a(i,.)\in[m,M]} W(i; a(i,.)). \tag{11}$$

The actions of this optimal sequence is equivalent to the optimal actions obtained by the belief-based value iteration given in (7), stated in the following proposition.

Proposition 1: There exist deterministic functions of s_L, last fully observed state, and t_L, time passed since observing the actual state, that determines the action selected by the optimal policy. In other words, the sequence achieving the supremum in (8) is equivalent to the sequence of the actions achieving the supremum in (6).

Proof: The solutions of the two value iterations given in (6) and (8) are equivalent since each pair of (s_L, t_L) corresponds to a specific PBD. The optimal policy for the belief-based formulation at each time step selects the action which achieves the supremum in (6) based on the PBD at that time step. For $t_L = 1$ passed from s_L, the PBD is equal to $f^{opt}_{s_L,1}(x) = p(x|s_L)$. Note that we use the subscript of s_L and t_L for PBD to show that this PBD corresponds to the case of passing t_L time steps from the last fully observed state s_L with no reset and we use the superscript *opt* for PBD to show that it is generated after selecting the optimal actions in the previous time steps. Now at the time step t_L, if we already know the optimal actions for the time steps $\tau = 1, 2, ..., t_L - 1$ passed from s_L, we can compute the corresponding PBD as follows:

$$f^{opt}_{s_L,\tau}(x) = \int_{m}^{M} T_{r^{opt}(s_L,\tau-1)}[f^{opt}_{s_L,\tau-1}](\alpha) p(x|\alpha) d\alpha,$$

for $x \in [m, M]$ and 0 otherwise and find the optimal action based on this PBD. Therefore, the optimal sequence found based on s_L and t_L corresponds to the optimal policy introduced in (7). Thus for any $s_L \in \mathcal{M}$,

$$r^{opt}(f^{opt}_{s_L,\tau}) = a^{opt}(s_L, \tau), \ \forall \tau = 1, 2, ...$$

□

V. Structural Properties of Myopic and Optimal Policies

In this section, we present some key properties of the myopic and optimal policies for both the belief-based and the sequence-based formulations. We show that any property which holds for the actions in the belief-based formulation is also valid for the sequences in the sequence-based formulation with some constraints on the transition probability densities.

A. Belief-Based Formulation: Properties of Myopic and Optimal actions

In the belief-based formulation, we can derive the myopic action which maximizes the immediate expected reward given in (4) and has a percentile threshold structure, for any PBD f, as follows.

$$r^{myopic}(f) = \inf\{r \in \mathcal{M} : \int_{x=m}^{r} f(x)dx = \frac{q+C_l}{q+C_l+C_u}\} = F^{-1}(\frac{q+C_l}{q+C_l+C_u}). \tag{12}$$

where $F^{-1}(y) = \inf_x\{F(x) \geq y\}$ is Inverse Cumulative Distribution Function (ICDF), and $F(x)$ is Cumulative Distribution Function (CDF) of the states. And the optimal action is bounded by the myopic action from below (See [2]),

$$r^{opt}(f) \geq r^{myopic}(f). \tag{13}$$

Now let us present an ordering of the myopic actions based on the ordering of PBDs defined below.

Definition 1: (First Order Stochastic Dominance, [12]) Let $f_1, f_2 \in \mathcal{B}$ be any two PBDs. Then f_1 First Order Stochastically dominates f_2 (or f_1 is FOSD greater than f_2), denoted as $f_1 \geq_s f_2$, if for all r, $F_1(r) \leq F_2(r)$ or equivalently,

$$\int_{x=r}^{\infty} f_1(x)dx \geq \int_{x=r}^{\infty} f_2(x)dx.$$

This ordering will be preserved for the updated PBD of the myopic policy at the next time step if the transition probability density has the FOSD-preserving property defined below.

Definition 2: (FOSD-preserving transition probability density) The transition probability density $p(x|y)$ is FOSD-preserving if for any $f_1 \geq_s f_2$,

$$\int_{y=m}^{M} f_1(y)p(x|y)dy \geq_s \int_{y=m}^{M} f_2(y)p(x|y)dy.$$

$$P^t_{i,a(i,1:t-1),j} \triangleq \int_{l_{t-1}=a(i,t-1)}^{M} \int_{l_{t-2}=a(i,t-2)}^{M} ... \int_{l_1=a(i,1)}^{M} p(l_1|i)...p(l_{t-1}|l_{t-2})p(l_j|l_{t-1})dl_1...dl_{t-2}dl'_{t-1} \quad (10)$$

The ordering of the myopic actions are given in the following lemma which is needed to prove the properties of the myopic and the optimal sequences in the next subsection.

Lemma 1: If f_1 and f_2 are two PBDs such that $f_1 \geq_s f_2$:

$$r_1^{myopic} \geq r_2^{myopic}, \quad (14)$$

$$T_{r_1^{myopic}}[f_1] \geq_s T_{r_2^{myopic}}[f_2], \quad (15)$$

where $r_i^{myopic} = r^{myopic}(f_i)$ is the myopic action corresponding to f_i for $i = 1, 2$.

Proof: Obviously, by definition of FOSD-ordering, the myopic actions for f_1 and f_2 obtained from (12) have the relationship given in (14). Now to prove (15) we have:

$$\int_{x=r}^{M} T_{r_1^{myopic}}[f_1](x)dx = \frac{\int_{x=r}^{M} f_1(x)dx}{\int_{x=r_1^{myopic}}^{M} f_1(x)dx}$$

$$\geq \frac{\int_{x=r}^{M} f_2(x)dx}{\int_{r_2^{myopic}}^{M} f_2(x)dx} = \int_{x=r}^{M} T_{r_2^{myopic}}[f_2](x)dx.$$

since $\int_{x=r_1^{myopic}}^{M} f_1(x)dx = \int_{r_2^{myopic}}^{M} f_2(x)dx$ and this completes the proof by Definition 2. □

Note that from (15) and FOSD-preserving property of the transition probability densities, the updated PBDs generated based on the previous myopic actions and also the corresponding myopic action will follow similar FOSD-orderings.

B. Sequence-Based Formulation: Properties of Myopic and Optimal sequences

In the sequence-based formulation, solving the value iteration to get the optimal action sequences is intractable. Instead, one simple sequence is the myopic sequence which can be derived from (12), $\forall i \in \mathcal{M}$, as follows:

$$a^{myopic}(i, t_L) = \inf\{r \in \mathcal{M} : \int_{j=m}^{r} P^{t_L}_{i,a^{myopic}(i,1:t_L),j} dj$$

$$= \frac{q + C_l}{q + C_l + C_u}\}, \ \forall t_L \geq 1.$$

Note that to compute the t_L-th action of the myopic sequence we should have computed the previous actions of the sequence. Now we present an ordering of the myopic sequences in the following proposition.

Proposition 2: For FOSD-preserving transition probability densities, we have the following properties for the myopic sequences with different last fully observed states:

$$a^{myopic}(i, t_L) \geq a^{myopic}(j, t_L), \ \forall i \geq j, \ \forall t_L,$$

which shows that the myopic sequence for the higher fully observed states is above the one for the lower fully observed states.

The proof could be achievable by induction on t_L and using Lemma 1. Now let us present the relationship between the optimal and the myopic sequences in the following theorem.

Theorem 1: The optimal sequence is lower bounded by the myopic sequence starting from the same fully observed state s_L.

$$a^{opt}(s_L, t_L) \geq a^{myopic}(s_L, t_L), \ \forall t_L.$$

The proof of the theorem is achievable using the following lemma.

Lemma 2: For FOSD-preserving transition probability densities, starting from the initial PBD f_0, the following relationships between the optimal and the myopic actions and the corresponding updated PBDs hold:

$$f_t^{opt} \geq_s f_t^{myopic}, \quad (16)$$

$$r_t^{opt} \geq r_t^{myopic}, \quad (17)$$

where f_t^{opt} and f_t^{myopic} are updated PBDs based on the optimal and myopic actions at previous time steps, r_τ^{opt} and r_τ^{opt} for $\tau = 1, 2, ..., t-1$, respectively.

Proof: We define a new set of actions $r_t^{m,o}$ which achieve the percentile threshold given in (12) on f_t^{opt}. Let us use induction to prove the above inequalities (16) and (17). To get (17) we will prove that:

$$r_t^{opt} \geq r_t^{m,o} \geq r_t^{myopic}. \quad (18)$$

The first inequality in (18) is achievable by (13). Now we use induction to prove (16) and the second inequality of (18). For the base of $t = 1$ by the assumption we have $f_1^{opt} = f_1^{myopic} = f_0$, this the second inequality in (18) for $t = 1$ holds as an equality.

Now by assuming they are valid for $t-1$, for t we get:

$$T_{r_{t-1}^{opt}} f_{t-1}^{opt} \geq_s T_{r_{t-1}^{m,o}} f_{t-1}^{opt} \geq_s T_{r_{t-1}^{myopic}} f_{t-1}^{myopic}. \quad (19)$$

The first inequality comes from the fact that $T_{r_1} f \geq_s T_{r_2} f, \forall r_1 \geq r_2$. The second inequality is achieved by (15), in Lemma 1. By applying FOSD-preserving transition probability densities to the PBDs in (19), we obtain (16), and from (14) we get the second inequality of (18). □

VI. Upper bound on Optimal Sequence

Beside the myopic sequence which provides a lower bound on the optimal sequence, we can derive a sequence as an upper-bound on the optimal sequence under zero under-utilization cost for a specific form of transition probability densities defined below.

Definition 3: (IIMC Process) The transition probability densities with the property of Independent Increment Markov Chain (IIMC) for the state space $\mathcal{M} = \mathbb{R}$, satisfies the following $\forall y \in \mathcal{M}$:

$$p(x|y) = p(x + \alpha|y + \alpha) \ \forall \alpha, x, y \in \mathbb{R}.$$

First we recall the following proposition which presents an upper bound on the optimal action from any PBD for our continuous-state problem. Later we will use this proposition to achieve the upper bound on the optimal sequences.

Proposition 3: (From [2]) For IIMC processes and $C_l = 0$, the optimal action is bounded from above by an action, denoted by r^{ub}, which is a function of β and the coefficients in the reward function, as follows:

$$r^{ub}(f) = F^{-1}(\frac{q+U}{q+C_u+U}), \tag{20}$$

where $U = \frac{q\beta}{1-\beta}(r^h - r^l)$ and $r^l = \sup\{x : f(x) \neq 0\}$ and $r^h = \inf\{x : f(x) \neq 0\}$ are the lowest and the highest states with non-zero probability densities, respectively.
The upper bound r^{ub} also has a percentile threshold structure with an extra term of U in the numerator and the denominator of the threshold.

Now let us present the upper bound on the optimal sequence which is achievable from the above proposition. For IIMC processes and $C_l = 0$, the upper-bound sequences denoted by $a^{UB}(i,.),\ \forall i \in \mathcal{M}$ are given by:

$$a^{UB}(i,t_L) = \inf\{r \in [r^l, r^h] : \int_{j=r^l}^{r} P^{t_L}_{i,a^{UB}(i,1:t_L),j} dj = \frac{q+U}{q+C_u+U}\},\ \forall t_L \geq 1.$$

where U is the same as what is defined in (20) and $r^h = \sup\{j : P^{t_L}_{i,a^{UB}(i,1:t_L),j} \neq 0\}$ and $r^l = \inf\{j : P^{t_L}_{i,a^{UB}(i,1:t_L),j} \neq 0\}$. Therefore we have the following theorem for the upper bound on the optimal sequence.

Theorem 2: The sequence a^{UB} is an upper bound on the optimal sequence, *i.e.* for any s_L,

$$a^{opt}(s_L,t_L) \leq a^{UB}(s_L,t_L),\ \forall t_L.$$

This upper bound sequence has the ordering property as follows.

Corollary 1: The sequence a^{UB} for FOSD-preserving transition probability densities, follows the monotonicity property with respect to the last fully observed state, *i.e.*,

$$a^{UB}(i,t_L) \geq a^{UB}(j,t_L),\ \forall i \geq j,\ \forall t_L.$$

We skip the proofs of the above theorem and corollary due to their similarities to Theorem 1 and Proposition 2.

VII. Summary and Conclusion

We have considered the tracking problem of real-valued Markovian random processes in which the goal is to select the best action sequences starting any full observation in order to get the supremum of the total expected discounted reward accumulated over an infinite horizon. We have modeled this decision-making problem as a POMDP in two different formulations and derived some key properties for the myopic and optimal policies.

We have shown that the actions can be defined with only two parameters: the last fully observed state and the time steps passed since the last observation. Therefore, we can present the optimal policy with the sequences starting from any fully observed state. We have proven that the whole optimal sequence is lower bounded by the myopic sequence starting from the same fully observed state.

We have presented some properties for myopic policy such as its percentile threshold structure, and its ordering under FOSD-preserving transition probability densities. Further, for IIMC processes, with zero under-utilization cost, we have derived an upper bound on the optimal sequence which also has a percentile threshold structure. As a future work, we will work on deriving the upper bound and an approximation for the optimal sequence for the general form of transition probability densities with similar percentile threshold structure.

References

[1] P. Mansourifard, B. Krishnamachari, and T. Javidi, "Bayesian congestion control over a markovian network bandwidth process," in *Signals, Systems and Computers, 2013 Asilomar Conference on*. IEEE, 2013, pp. 332–336.

[2] P. Mansourifard, T. Javidi, and B. Krishnamachariy, "Tracking of markovian random processes with asymmetric cost and observation," *available online http://anrg.usc.edu/www/papers/TMRPACO-submitted.pdf*, 2014.

[3] A. Laourine and L. Tong, "Betting on gilbert-elliot channels," *Wireless Communications, IEEE Transactions on*, vol. 9, no. 2, pp. 723–733, 2010.

[4] L. A. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A pomdp search theory formulation with optimal threshold policies," *Wireless Communications, IEEE Transactions on*, vol. 5, no. 2, pp. 394–405, 2006.

[5] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of markov decision processes," *Mathematics of operations research*, vol. 12, no. 3, pp. 441–450, 1987.

[6] Y. Wu and B. Krishnamachari, "Online learning to optimize transmission over an unknown gilbert-elliott channel," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on*. IEEE, 2012, pp. 27–32.

[7] Y. Qin, R. Wang, A. J. Vakharia, Y. Chen, and M. M. Seref, "The newsvendor problem: Review and directions for future research," *European Journal of Operational Research*, vol. 213, no. 2, pp. 361–374, 2011.

[8] X. Ding, M. L. Puterman, and A. Bisi, "The censored newsvendor and the optimal acquisition of information," *Operations Research*, vol. 50, no. 3, pp. 517–527, 2002.

[9] O. Besbes and A. Muharremoglu, "On implications of demand censoring in the newsvendor problem," *Management Science, Forthcoming*, pp. 12–7, 2010.

[10] A. Bensoussan, M. Çakanyıldırım, and S. P. Sethi, "A multiperiod newsvendor problem with partially observed demand," *Mathematics of Operations Research*, vol. 32, no. 2, pp. 322–344, 2007.

[11] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[12] A. Muller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*. Hoboken, NJ: Wiley, 2002.