Ming Hsieh Department of Electrical Engineering

USC Viterbi School of Engineering

ANRG

Sequence Based Tracking of Continuous Markovian Random Processes with Asymmetric Cost and Observation

Parisa Mansourifard, Bhaskar Krishnamachari , Tara Javidi

Introduction

- We consider a state-tracking problem in which the background random process is Markovian with continuous states.
- At each time step the decision-maker chooses as an action a state and accumulates some reward based on the selected state and the actual state.
- The goal is to select the actions such that the total expected discounted reward is maximized.
- We model this problem as a Partially Observable Markov Decision Process and formulate it in two different ways:
  (i) belief-based value function,
  (ii) sequence-based value function.
- In the sequence-based formulation, only two parameters matter to define the sequences, the last observed state and the time passed from the last observation.

POMDP

- **State:** The actual state of the Markov process B_t at time step t, can be any real number the state space, i.e. $[m, M]$

- **State transition:** The transition probabilities of the actual states over time are shown by

$$p(x \mid y) := P(B_t = x \mid B_{t-1} = y), \forall m \le x, y \le M$$

- **Action:** At each time step, we choose an action r_t from the action space which is equivalent to the state space.

- **Observed information:** The observed information at time step t is defined by the event $o_t(r_t)$
The possible events corresponding to the action r_t

$$o_t(r_t) = \{B_t = i\}, \forall i \in [m, r_t)$$

the event of fully observing the actual state.

$$o_t(r_t) = \{B_t \ge r_t\}$$

the event of partial observing that the actual state is larger than or equal to the selected state.

- **Reward:** The immediate reward earned at time step t is defined as follows:

$$R(B_t, r_t) = \begin{cases} q\,B_t - C_u(r_t - B_t), & \text{if } r_t > B_t \\ q\,r_t - C_l(B_t - r_t), & \text{if } r_t \le B_t \end{cases}$$

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

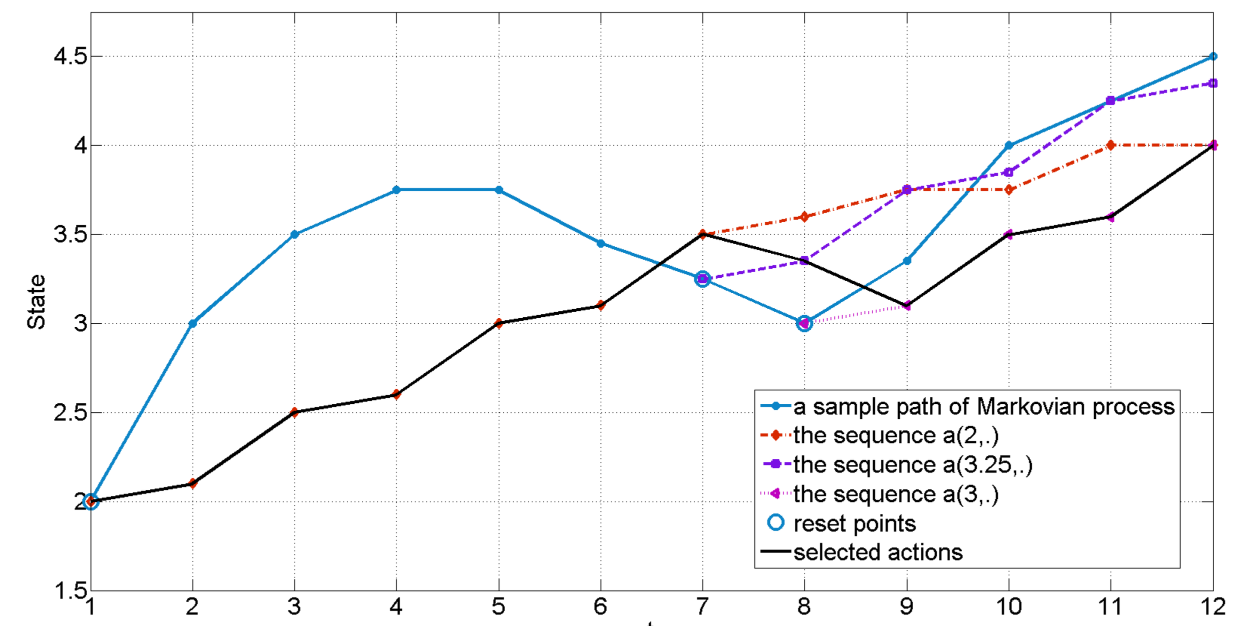
Sequence-Based Formulation

decision-maker decides about the whole sequence after any full observation about the actual state.
The optimal policy can also be perfectly characterized by only two parameters;
(i) the last observed state, s_L
(ii) the time steps passed since the last observation, t_L

The sequence of actions starting from state s_L

$$a(s_L, .) = \{a(s_L, 1), a(s_L, 2), \ldots\}$$

An Example of Action Sequences



Myopic Policy

Myopic policy maximizes the immediate expected reward ignoring the future

$$a^{myopic}(s_L, t_L) = \inf\ \{r \in [m, M]:$$

$$\int_{j=m}^{r} P^{t_L}_{i, a^{myopic}(s_L, 1:t_L), j}\, dj = \frac{q + C_l}{q + C_l + C_u}\}$$

Where $P^{t_L}_{i, a^{myopic}(s_L, 1:t_L), j}$ indicates the probability of occurring the following event: no reset {i.e. full observation) at time steps $1, 2, \ldots, t_L - 1$ passed from the last observed state s_L and following the action sequence of $a(s_L, 1:t_L) = \{a(s_L, 1), \ldots, a(s_L, t_L)\}$ and reset to the actual state j at t_L.

Main Theorem

- The optimal sequence is lower bounded by the myopic sequence started from the same observed state, i.e.

$$a^{opt}(s_L, t_L) \ge a^{myopic}(s_L, t_L)$$

parisama@usc.edu, tjavidi@ucsd.edu, bkrishna@usc.edu