Bayesian Congestion Control over a Markovian Network Bandwidth Process

Parisa Mansourifard,[†], Bhaskar Krishnamachari[†], Tara Javidi[‡]
[†] Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA
[‡] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA
emails: parisama@usc.edu, bkrishna@usc.edu, tjavidi@ucsd.edu

Abstract—**We formulate a Bayesian congestion control problem in which a source must select the transmission rate over a network whose available bandwidth is modeled as a time-homogeneous finite-state Markov Chain. The decision to transmit at a rate below the instantaneous available bandwidth results in an under-utilization of the resource while transmission at rates higher than the available bandwidth results in a linear penalty. The trade-off is further complicated by the asymmetry in the information acquisition process: transmission rates that happen to be larger than the instantaneous available bandwidth result in perfect observation of the state of the bandwidth process. In contrast, when transmission rate is below the instantaneous available bandwidth, only a (potentially rather loose) lower bound on the available bandwidth is revealed. We show that the problem of maximizing the throughput of the source while avoiding congestion loss can be expressed as a Partially Observable Markov Decision Process (POMDP). We prove structural results providing bounds on the optimal actions. The obtained bounds yield tractable sub-optimal solutions that are shown via simulations to perform well.**

I. Introduction

In many network protocols, a device 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.

In our congestion control problem, we consider a user whose available bandwidth varies as a Markovian process. A sender wants to set its transmission rate at each time step. If the sender selects a rate higher than the available bandwidth, it can utilize the whole available bandwidth but has to pay an over-utilization penalty (which is a function of how much the selected rate exceeds the available bandwidth). In this case, we assume that perfect information about the current available bandwidth is revealed (full observation). If the user selects a rate less than the available bandwidth, it does not experience loss (no penalty), but the available bandwidth is under-utilized. Also, in this case, the sender gets partial information about the available bandwidth, *i.e.* it can infer that the current available bandwidth is larger than or equal to the selected rate, but not its exact value. In such a setting, there is a trade-off between getting more information about the available bandwidth and paying less penalty. The goal is to find the optimal policy to maximize the total reward (utilization minus penalty) over a finite horizon.[1]

We formalize this problem as a Partially Observable Markov Decision Process (POMDP) [1], since the state of underlying Markov chain, called the "resource state", is not fully observed. Since the POMDP problem does not have an efficiently computable solution [2], we instead present upper and lower bounds on the optimal actions. The lower bounds correspond to the myopic policy which at each time step selects an action maximizing the immediate expected reward, ignoring the future. Intuitively, selecting actions higher than the myopic action results in less immediate reward but also more information about the resource state. Thus, to learn more about the resource state the optimal action prefers to exceed the myopic action. We prove that the myopic policy has a percentile threshold structure, *i.e.* it selects an action corresponding to the lowest state above a given percentile of the beliefs. The upper bound on the optimal action has a similar closed-form structure, based on the belief vector and the system parameters. In order to derive the upper bound, we further assume that the background Markov chain has a particular State-Independent State Change (SISC) structure (see Section III for more details). We consider the effect of different parameters on the bounds via simulation.

The remainder of the paper is organized as follows: Related works are presented in Section II. In Section III, the problem formulation is introduced. In Section IV, the main results of the paper are derived. Section V presents some key properties needed for deriving our main results. Simulation results are presented in Section VI. Finally, Section VII concludes the paper.

II. Related Work

Congestion control is a fundamental problem in networking and has a rich literature on algorithms and theoretical analyses (see *e.g.* [3], [4]). Existing techniques, such as Transmission Control Protocol (TCP), adopt an Additive Increase Multiplicative Decrease (AIMD) algorithm, introduced in [5], that adjusts the congestion window based on the transmission acknowledgments. The performance of AIMD is analyzed in literature, *e.g.* [6] where Altman *et al.* introduce a general model based on a multi-state Markov chain for the moments at which the congestion is detected. Although AIMD often

[1] The results in this paper can be readily extended to infinite horizon discounted reward

works well in practice, it is a heuristic that is not guaranteed to be optimal. In this work, we aim to identify the structure of optimal congestion control, albeit under the special case of a Markovian available bandwidth process.

Threshold structures of optimal transmission policies have been established for simpler related problems with a two-state Gilbert-Elliott channel in [7], [8]. But here, we consider a more general case of multi-state Markovian channel

Some recent works in literature, *e.g.*, [9], [10], formulate the problem of inventory management where some number of items needs to be stock (inventory) and the demand is a stochastic process. The problem of inventory management has a close connection to our problem as we can map the demand to the resource (bandwidth) and the inventory level to the action (rate allocated). Of these, the most closely related work is that of Bensoussan *et al.* [9], who consider a POMDP problem where the demand is a Markovian process. They consider the setting where the resources and actions are both continuous, as well as the case where the resources are discrete but the actions remain continuous. They use the un-normalized beliefs to prove the existence of the optimal policy which is challenging for the continuous setting. For these settings, they also show that the optimal actions exceed the myopic actions. In this paper, in contrast to their work, we consider the case where both the resource states and the actions are discrete and finite. Thus, the existence of the optimal policy is trivial [1]. Further, by investigating a specific form of the transition probability, SISC, we derive additional properties of the optimal policy and the reward functions. Furthermore, this work is the first to present an upper bound for the optimal action. Besides giving insight into the optimal actions, this bound can also be used to speed up the search for the optimal actions in the dynamic programming.

In [10], Besbes *et al.* consider a similar problem with the assumption that the resource state is an independent-identical distribution process which is a specific form of Markovian processes. They assume that the underlying resource distribution is not available and need to be estimated from historical data. They show that a percentile threshold policy (similar to our myopic action) is optimal and characterize the implications of partial observations on the performance of the optimal policy in both discrete and continuous settings. We consider the Markovian case in which the given percentile threshold policy constructs a lower bound for the optimal actions.

III. Problem Formulation

We consider a discrete-time finite-state Markov process, whose state is denoted by B_t. At each time step, the user selects an action according to the history of observations- thus earning a reward as a function of the selected action and the state B_t and gathers some partial information about the current state. The objective is to find an optimal policy as a sequence of actions to select in order to maximize the total expected discounted reward accumulated over the finite horizon.

Let us denote the finite horizon by T and let the discrete time steps be indexed by $t = 1, 2, ..., T$. We formulate our problem within a POMDP-based framework defined as a tuple $\{\mathcal{M}, P, \mathcal{B}, \mathcal{A}, \mathcal{O}, U, R\}$ where:

- *State:* The state, B_t, is one of the elements of a finite set denoted by $\mathcal{M} = \{1, 2, ..., M\}$
- *State transition:* The state B_t varies over time according to a Markov process with a known transition probability matrix, denoted by P. This matrix is an $M \times M$ matrix with elements $P_{i,j} = Pr(B_{t+1} = j | B_t = i), i, j \in \mathcal{M}$ which indicates the probability of moving from state i at a time step to the state j at the next time step.
- *Belief vector:* The probability distribution of the resource state (assuming a finite state set), given all past observations, is denoted by a belief vector $b_t = [b_t(1), ..., b_t(M)]$, with elements of $b_t(k) = Pr(B_t = k), k \in \mathcal{M}$. In other words, b_t represents the probability distribution of B_t over all possible states of $\mathcal{M}$. The set of all possible belief vectors is denoted by $\mathcal{B}$. The goal is to make a decision at each time step based on the history of observations; but due to the lack of full information, the decision should be based on the belief vector. It can be shown that the belief vector is a sufficient statistic of the complete observation history (see *e.g.*, [1]).
- *Action:* At each time step, according to the current belief, we choose an action $r_t \in \mathcal{A} = \{1, ..., M\}$. Note the set of actions are equal to the set of states, *i.e.* $\mathcal{A} = \mathcal{M}$.
- *Observed information:* The observed information at time step t is defined by the event $o_t(r_t) \in \mathcal{O}$ as a function of selected action. The possible events corresponding to the action r_t is as follows:
  - $o_t(r_t) = \{B_t = i\}, i = 1, ..., r_t - 1$ is the event of fully observing B_t. This corresponds to the selection of the action higher than B_t.
  - $o_t(r_t) = \{B_t \geq r_t\}$ is the event that B_t is larger than or equal to the selected action.
- *Belief updating:* The belief updating is a mapping $U : \mathcal{A} \times \mathcal{O} \times \mathcal{B} \to \mathcal{B}$. The belief vector for the next time step, upon the selected action and the observation, is governed by:

  $$b_{t+1} = \begin{cases} T_{r_t} b_t P & \text{if } r_t \leq B_t \\ I_{B_t} P & \text{if } r_t > B_t, \end{cases} \quad (1)$$

  where I_a is the M-dimensional unit vector with 1 in the a-th position. T_r is a non-linear operation on a belief vector b, as follows:

  $$T_r b(i) = \begin{cases} 0 & \text{if } i < r \\ \frac{b(i)}{\sum_{j=r}^{M} b(j)} & \text{if } i \geq r. \end{cases} \quad (2)$$

- *Reward:* The immediate reward earned at time step t is a mapping $R : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$, which is given by:

  $$R(B_t, r_t) = \begin{cases} B_t - C(r_t - B_t) & \text{if } r_t > B_t \\ r_t & \text{if } r_t \leq B_t, \end{cases} \quad (3)$$

  where C is the over-utilization penalty coefficient.

The policy π specifies a sequence of functions $\pi_1, ..., \pi_T$, where π_t is the decision rule and maps a belief vector b_t to an action at time step t, *i.e.*, $\pi_t : \mathcal{B} \to \mathcal{A}$, $r_t = \pi_t(b_t)$. The goal is to maximize the total expected discounted reward in finite horizon, over all admissible policies π, given by

$$\max_{\pi} J_T^{\pi}(b_0) = \max_{\pi} E^{\pi}[\sum_{t=0}^{T} \beta^t R(B_t; r_t)|b_0], \tag{4}$$

where $0 \leq \beta \leq 1$ denotes the discount factor and b_0 is the initial belief vector. $J_T^{\pi}(b_0)$ is the total expected discounted reward accumulated over the horizon T under policy π and starting in the initial belief vector b_0. The optimal policy denoted by π^{opt} is a policy which maximizes (4) and it exists since the number of admissible policies are finite.

We may solve this POMDP problem using Dynamic programming (DP), as the following recursive equations holds:

$$V_t(b_t) = \max_{r_t} V_t(b_t; r_t), \tag{5a}$$

$$V_t(b_t; r_t) = \bar{R}(b_t; r_t) + \beta E\{V_{t+1}(b_{t+1})|r_t\}, \ \forall t \neq T \tag{5b}$$

$$V_T(b_T; r_T) = \bar{R}_T(b_T; r_T), \tag{5c}$$

where b_{t+1} is the updated belief vector for the next time step shown in (1). The value functions $V_t(b_t)$ is the maximum remaining expected reward accrued starting from time t when the current belief vector is b_t. Note for all $t = 1, ..., T$, $V_t(b_t) = \max_{\bar{r}} J_{T-t+1}^{\pi}(b_t)$. $V_t(b_t; r_t)$ is the remaining expected reward accrued after time t with choosing action r_t at time t and following the optimal policy for time $t+1, ..., T$ with updated belief vector according to the action r_t. $V_t(b_t; r_t)$ is the summation of two terms: (i) the immediate expected reward, given by taking expectation of (3):

$$\begin{aligned} \bar{R}(b_t; r_t) &= \sum_{i \in \mathcal{M}} b_t(i) R(i, r_t) \\ &= r_t \sum_{i=r_t}^{M} b_t(i) + \sum_{i=1}^{r_t - 1} b_t(i)[(1+C)i - C r_t], \end{aligned} \tag{6}$$

and (ii) the discounted future expected reward which can be computed as follows:

$$\begin{aligned} V_t^f(b_t; r_t) &= E\{V_{t+1}(b_{t+1})|r_t\} \\ &= \sum_{i=r_t}^{M} b_t(i) V_{t+1}(T_{r_t} b_t P) + \sum_{i=1}^{r_t - 1} b_t(i) V_{t+1}(I_i P). \end{aligned} \tag{7}$$

A policy π is optimal if and only if for $t = 1, ..., T$, $r_t = \pi_t(b_t)$ achieves the maximum in (5a), denoted by:

$$r_t^{opt}(b) = \arg\max_{r \in \mathcal{A}} V_t(b; r). \tag{8}$$

We present upper and lower bounds on the optimal actions in two theorems given in the next section using some lemmas. All the proofs are given in the appendices. For some of our results we need the following assumptions.

Assumption 1. The P matrix satisfies the State-Independent State Change (SISC) property.

Assumption 2. The P matrix satisfies the State-Independent State Change (SISC) property with edge effects, defined below.

By SISC property, we mean that $P_{i,i+k} = p_{j,j+k}$. Therefore, we can define our transition matrix P by indicating the probability of moving k step higher, p_k, independent of which state we are, such that $k < 0$ corresponds to moving $-k$ steps lower.

By edge effect, we mean that the transition matrix will be affected by the limits (edges) of the state set, since the state set $\mathcal{M}$ is limited from both sides. Therefore, the elements of the P matrix for our desired SISC process are equal to $P_{i,j} = p_{j-i}$, for all i, j except the following:

$$P_{1,j} = P_{1,j+1} + P_{2,j+1}, \qquad j \leq M - 1 \tag{9}$$

$$P_{M,j} = P_{M,j-1} + P_{M-1,j-1}, \qquad j \geq 2 \tag{10}$$

where (9) and (10) reflect the lower and upper edge effects of $\mathcal{M}$, respectively.

Note that for simplicity of notations, from now on we drop the subscripts of t from the belief vectors and the actions.

IV. Main Results

The main result of our paper is stated in the following two theorems.

Theorem 1. The optimal action is bounded by an action from below, denoted by r^{lb}, given by

$$r^{lb} = \min\{r \in \mathcal{M} : \sum_{i=1}^{r} b(i) \geq \frac{1}{1+C}\}. \tag{11}$$

This lower bound is equal to the myopic action which at each time step selects the action maximizing the immediate expected reward and ignores its impact on the future reward. The myopic action, for problem (4), under belief vector b is given by:

$$r^{myopic}(b) = \arg\max_{r \in \mathcal{M}} \bar{R}(b; r). \tag{12}$$

Theorem 2. Under Assumption 1 or 2, the optimal action is bounded from above by an action, denoted by r^{ub}, which is a function of C, β, and belief vector b as follows:

$$\begin{aligned} r^{ub} = &\min\{r \in \mathcal{M} : \\ &f(\beta)\bar{S}_r + [(1+C) - f(\beta)r]S_r - C \leq 0\}, \end{aligned} \tag{13}$$

where for simplicity we define the following notations:

$$S_r \triangleq \sum_{i=r+1}^{r^h} b(i), \quad \bar{S}_r \triangleq \sum_{i=r+1}^{r^h} i b(i), \quad f(\beta) \triangleq \beta \frac{1-\beta^{T-1}}{1-\beta}. \tag{14}$$

Note that from the above inequalities, $r^l \leq r^{lb} = r^{myopic} \leq r^{opt} \leq r^{ub} \leq r^h$, where r^l and r^h are the lowest and the highest states with non-zero beliefs, respectively.

V. Analyses

To prove Theorem 1, we need the following lemmas.

Lemma 1. The expected immediate reward is a uni-modal function of the action r, which is increasing when $\sum_{i=1}^{r} b(i) < \frac{1}{1+C}$ and it is decreasing otherwise. Therefore, the myopic action which maximizes the expected immediate reward (given in (12)) is equal to the lowest action r satisfying the inequality $\sum_{i=1}^{r} b(i) \geq \frac{1}{1+C}$. Interestingly, the myopic action has a percentile threshold structure, *i.e.* it corresponds to the lowest state above a given percentile of the beliefs.

Lemma 2. The remaining expected discounted reward in the action r, $V_t(b;r)$, and the value function, $V_t(b)$, are convex with respect to the belief vector b, *i.e.*

$$V_t(b;r) \leq \lambda V_t(b_1;r) + (1-\lambda)V_t(b_2;r), \quad \forall r \in \mathcal{M},$$
$$V_t(b) \leq \lambda V_t(b_1) + (1-\lambda)V_t(b_2), \quad \forall 0 \leq \lambda \leq 1. \tag{15}$$

Lemma 3. The future expected reward, $V_t^f(b;r)$ defined in (7), is monotonically increasing in the action, *i.e.*,

$$V_t^f(b;r_1) - V_t^f(b;r_2) \geq 0, \ \forall r_1 \geq r_2. \tag{16}$$

Now let define an ordering for the belief vectors and derive a key property of value functions, their monotonicity with respect to an ordering of the belief vectors. Then we state some lemmas which are needed to prove Theorem 2.

Definition 1. (First Order Stochastically Dominance, [11]) Let $b_1, b_2 \in \mathcal{B}$ be any two belief vectors. Then b_1 first order stochastically dominates b_2 (or b_1 is FOSD greater than b_2), denoted as $b_1 \geq_s b_2$, if

$$\sum_{j=r}^{M} b_1(j) \geq \sum_{j=r}^{M} b_2(j), \ r \in \{1, ..., M\}. \tag{17}$$

Lemma 4. Under Assumption 1 or 2, the value function is a FOSD-increasing function of the belief vector. *i.e.*, if $b_1 \geq_s b_2$, then $V(b_1) \geq V(b_2)$.

Now let consider two belief vectors b and b^α such that b^α is shifted version of b by α steps, *i.e.* $b^\alpha(i) = b(i+\alpha)$. Because of similarity in their probability distribution shapes, they have some common properties, given in the following lemmas.

Lemma 5. The immediate expected reward and the myopic action follow the shifting property:

$$\bar{R}(b^\alpha;r) = \bar{R}(b;r-\alpha) + \alpha, \tag{18}$$
$$r^{myopic}(b^\alpha) = r^{myopic}(b) + \alpha. \tag{19}$$

Lemma 6. Under Assumption 1, the value functions and the optimal policies of the belief vector b and its shifted version, b^α, have the following properties:

$$V_t(b^\alpha;r) - V_t(b;r-\alpha) = \alpha\frac{1-\beta^{T-t+1}}{1-\beta}, \tag{20}$$
$$r_t^{opt}(b^\alpha) = r_t^{opt}(b) + \alpha, \tag{21}$$
$$V_t(b^\alpha) = V_t(b) + \alpha\frac{1-\beta^{T-t}}{1-\beta}. \tag{22}$$

Note for $\beta = 1$, we need to substitute $\frac{1-\beta^x}{1-\beta}$ by x.

Lemma 7. Under Assumption 2, the following relation between the value functions of b and b^α holds:

$$V_t(b^\alpha) \leq V_t(b) + \alpha\frac{1-\beta^{T-t}}{1-\beta}. \tag{23}$$

The proofs of the above lemmas are given in Appendix A, followed by the proofs of Theorems 1 and 2 in Appendix B.

VI. Simulation

We present some simulation results to consider the upper and the lower bounds achieved in the previous section. The simulation parameters, except in the figures that their effect is considered, are fixed as follows: the number of states $M = 10$, the over-utilization penalty coefficient $C = 5$, the discount factor $\beta = 0.8$, and the transition probabilities $p_1 = p_{-1} = 0.3, p_0 = 0.4$. Due to computational complexity it is hard to find the optimal policy. Instead, we could consider the Extended Myopic (EM) policy with future window of size τ as an approximation for the optimal policy. This policy, at each time step t, solves the dynamic programming for short horizon of $t, t+1, ..., t+\tau$. Here we use $\tau = 4$ for our simulations. Note that we could limit our dynamic programming searches between the upper and the lower bound to speed up the simulations.

A. Upper and Lower Bound on Optimal Actions

Fig. 1 shows an example of a sequence of optimal actions and their corresponding upper and lower bounds. We assume that the selected actions does not exceed B_t till time step 14. Note that the stars in the figure indicate the non-zero beliefs.

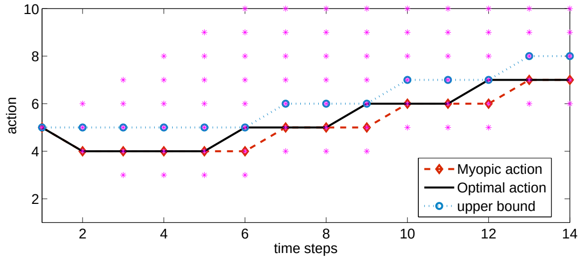

Fig. 1. Selected actions by optimal policy (EM, $\tau = 4$) and their corresponding lower and upper bounds, for $C = 5$, $\beta = 0.8$, and transition of $p_1 = p_{-1} = 0.3, p_0 = 0.4$

Next, we consider the effect of β on the gap between the upper and the lower bounds of optimal actions in Fig. 2. As expected, this gap increases with increasing β.

Fig. 3 shows the gap between the upper and the lower bounds versus the variance of SISC transition, defined as $E[(B_{t+1} - B_t)^2]$. This figure confirms that by increasing the variance, the gap will increase.

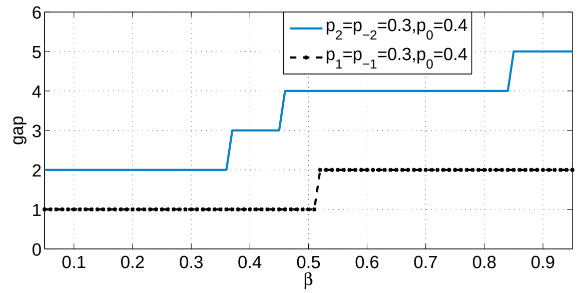

Fig. 2. The gap between the lower and the upper bounds versus β, for $C = 5$.

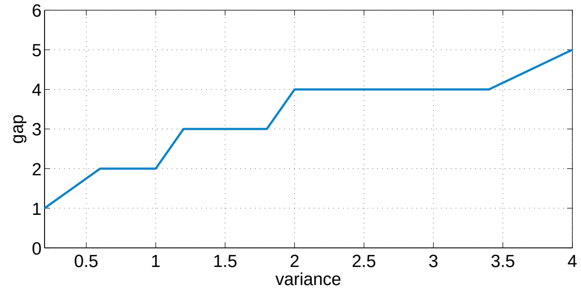

Fig. 3. The gap between the lower and the upper bounds versus the variance of random walk steps, for $\beta = 0.8$, $C = 5$.

B. Myopic and Upper-Bound policies

Now we compare two sub-optimal policies: (i) the myopic policy, (ii) the upper-bound (UB) policy, which pick the actions given in (11) and (13), respectively, at all time steps and update their belief vectors according to these actions. Fig. 4 shows the total expected discounted reward versus β for the myopic and the UB policies. For smaller β the reward of the myopic policy is higher than that of UB policy, but for larger β (close to 1) the UB policy outperforms the myopic policy. Fig. 5 shows the total expected discounted reward versus C for the myopic and the UB policies for $\beta = 0.8$. The difference between the reward of two policies will increase as C grow.

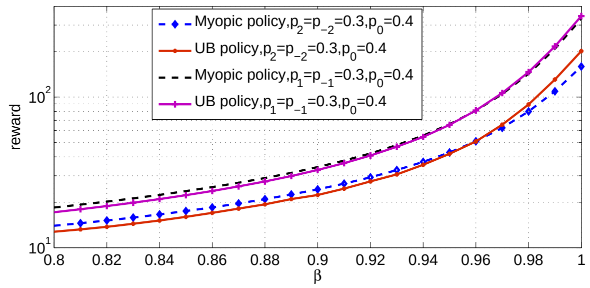

Fig. 4. The total expected discounted reward for two sub-optimal policies versus β for $C = 5$, $\tau = 4$, for horizon $T = 100$.

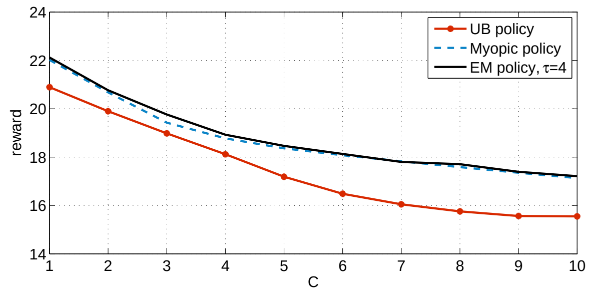

Fig. 5. The total expected discounted reward for two sub-optimal policies versus C for $\beta = 0.8$, $\tau = 4$, for horizon $T = 100$ and transition of $p_1 = p_{-1} = 0.3, p_0 = 0.4$.

VII. Conclusion

We formulated a Bayesian congestion control problem in which a source must select the transmission rate (the action) over a network whose available bandwidth (resource) evolves as a stochastic process. We modeled the problem as a POMDP and derived some key properties for the myopic and the optimal policies. We proved structural results providing bounds on the optimal actions, yielding tractable sub-optimal solutions that have been shown via simulations to perform well. Since the myopic action has a percentile threshold structure on the state beliefs, and we can simplify our upper bound to get a looser one with a similar form, we conjecture that there may be even better approximation for the optimal policy with the similar percentile threshold structure.

Acknowledgement

This work was supported in part by the U.S. National Science foundation under ECCS-EARS awards numbered 1247995 and 1248017 and by the Okawa foundation through an award to support research on ”Network Protocols that Learn”.

References

[1] 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.

[2] 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.

[3] R. Srikant, *The mathematics of Internet congestion control*. Birkhauser Boston, 2004.

[4] C. Jin, D. X. Wei, and S. H. Low, “Fast tcp: motivation, architecture, algorithms, performance,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 4. IEEE, 2004, pp. 2490–2501.

[5] V. Jacobson, “Congestion avoidance and control,” in *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4. ACM, 1988, pp. 314–329.

[6] E. Altman, K. Avrachenkov, C. Barakat, and P. Dube, “Performance analysis of aimd mechanisms over a multi-state markovian path,” *Computer Networks*, vol. 47, no. 3, pp. 307–326, 2005.

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

[8] 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.

[9] 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.

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

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

Appendix A

Proof of Lemma 1:

To prove this lemma, let us compute the derivative of the expected immediate reward given in (6),

$$\begin{aligned}\Delta\bar{R}(b;r) &= \bar{R}(b;r+1) - \bar{R}(b;r)\\ &= 1 - (1+C)\sum_{i=1}^{r} b(i). \end{aligned} \tag{24}$$

It’s easily seen that if the inequality in (11) holds, $\Delta\bar{R}(b;r) \leq 0$. Otherwise it is positive. This concludes that $\bar{R}(b;r)$ is unimodal with unique maximum at the myopic action given in (11). ■

Proof of Lemma 2:

We use induction to prove the convexity of $V_t(b;r)$ with respect to the belief vector, b, for finite horizon. Let’s assume b is a linear combination of two belief vectors b_1 and b_2, such that:

$$b = \lambda b_1 + (1-\lambda)b_2, 0 \leq \lambda \leq 1. \tag{25}$$

First at horizon T, the remaining expected reward is equal to the expected immediate reward which is affine linear with respect to the belief vector and using (6) and (25) we have:

$$\bar{R}(b;r) = \lambda\bar{R}(b_1;r) + (1-\lambda)\bar{R}(b_2;r), \tag{26}$$

which confirms the convexity of the remaining expected reward at horizon T. Now assuming the convexity holds at $t+1$, we will consider the time step t. Using (5b) and (7) we have:

$$\begin{aligned} &V_t(b;r) - \lambda V_t(b_1;r) - (1-\lambda)V_t(b_2;r)\\ &= [R(b;r) - \lambda\bar{R}(b_1;r) - (1-\lambda)\bar{R}(b_2;r)]\\ &+ \beta[V_{t+1}(T_r bP)\sum_{i=r}^{M} b(i) - \lambda V_{t+1}(T_r b_1 P)\sum_{i=r}^{M} b_1(i)\\ &- (1-\lambda)V_{t+1}(T_r b_2 P)\sum_{i=r}^{M} b_2(i)] && \text{(27a)}\\ &= \beta\sum_{i=r}^{M} b(i)[V_{t+1}(T_r bP) - \lambda' V_{t+1}(T_r b_1 P)\\ &- (1-\lambda')V_{t+1}(T_r b_2 P)] && \text{(27b)} \end{aligned}$$

where $\lambda' = \lambda\frac{\sum_{i=r}^{M} b_1(i)}{\sum_{i=r}^{M} b(i)}$. Note using (26), the term inside [.] in (27a) is zero. Multiplying the transition matrix P is a linear operation, and with some manipulation, we can show that $\lambda' T_r b_1 + (1-\lambda')T_r b_2 = T_r b$:

$$\begin{aligned} A &= \lambda' T_r b_1 + (1-\lambda')T_r b_2\\ &= \frac{\lambda\sum_{i=r}^{M} b_1(i)T_r b_1 + (1-\lambda)\sum_{i=r}^{M} b_2(i)T_r b_2}{\sum_{i=r}^{M} b(i)}\\ A(j) &= \frac{1}{\sum_{i=r}^{M} b(i)}[\lambda\sum_{i=r}^{M} b_1(i)\frac{b_1(j)}{\sum_{i=r}^{M} b_1(i)}\\ &+ (1-\lambda)\sum_{i=r}^{M} b_2(i)\frac{b_2(j)}{\sum_{i=r}^{M} b_2(i)}]\\ &= \frac{\lambda b_1(j) + (1-\lambda)b_2(j)}{\sum_{i=r}^{M} b(i)} = \frac{b(j)}{\sum_{i=r}^{M} b(i)}, \quad j \geq r\\ A(j) &= 0, \quad j < r \end{aligned} \tag{28}$$

So from definition of (2), we have $A = T_r b$. Now from convexity at $t+1$ the term inside [.] in (27b) is less than or equal to zero and we have:

$$V_t(b;r) - \lambda V_t(b_1;r) - (1-\lambda)V_t(b_2;r) \leq 0 \tag{29}$$

which means the convexity of $V_t(b;r)$ with respect to b.

To prove the convexity of value function, $V_t(b)$, with respect to b we use the definition of (5a) and the result of Lemma 2 to get:

$$\begin{aligned} V_t(b) &= \max_{r} V_t(b;r) = V_t(b;r^*) && \text{(30a)}\\ &\leq \lambda V_t(b_1;r^*) + (1-\lambda)V_t(b_2;r^*) && \text{(30b)}\\ &\leq \lambda\max_{r_1} V_t(b_1;r_1) + (1-\lambda)\max_{r_2} V_t(b_2;r_2) && \text{(30c)}\\ &= \lambda V_t(b_1) + (1-\lambda)V_t(b_2), && \text{(30d)} \end{aligned}$$

where $r^* = \arg\max_r V_t(b;r)$. Applying the definition of (5a) one more time in (30d) completes the proof. ■

Proof of Lemma 3:

We can write (16) as follows:

$$\begin{aligned} &V_t^f(b;r_1) - V_t^f(b;r_2)\\ &\qquad = \sum_{r=r_2}^{r=r_1-1}\{V_t^f(b;r+1) - V_t^f(b;r)\} \end{aligned} \tag{31}$$

and for each term inside the summation, using (7), we have:

$$\begin{aligned} &\Delta V_t^f(b;r) = V_t^f(b;r+1) - V_t^(b;r)\\ &= \sum_{i=r+1}^{M} b(i)V_{t+1}(T_{r+1}bP) - \sum_{i=r}^{M} b(i)V_{t+1}(T_r bP)\\ &+ b(r)V_{t+1}(I_r P) \geq 0 \end{aligned} \tag{32}$$

The inequality (32) is achieved from the convexity of value function in (15) for $b = T_r bP$, $b_1 = T_{r+1}bP$, $b_2 = I_r P$ and $\lambda = \frac{\sum_{i=r+1}^{M} b(i)}{\sum_{i=r}^{M} b(i)}$. Therefore, (31) is greater than or equal to zero and proof is complete. ■

Proof of Lemma 4:

If $b_1 \geq_s b_2$, *i.e.*, by recalling Definition 1,

$$\sum_{j=r}^{M} b_1(j) \geq \sum_{j=r}^{M} b_2(j),\ \ r \in \{1, ..., M\}, \tag{33}$$

we can write b_1 as a linear combination of some belief vectors with coefficients equal to the elements of b_2, as follows:

$$\begin{aligned} b_1 &= \sum_{i=1}^{j} b_2(i) b'_i, \\ b'_i &\in \mathcal{B}, b'_i(j) = 0, j < i. \end{aligned} \tag{34}$$

We can obtain (33) using (34), as follows:

$$\begin{aligned} \sum_{j=r}^{M} b_1(j) &= \sum_{j=r}^{M} \sum_{i=1}^{j} b_2(i) b'_i(j) \\ &\geq \sum_{j=r}^{M} \sum_{i=r}^{j} b_2(i) b'_i(j) && \text{(35a)} \\ &\geq \sum_{i=r}^{M} \sum_{j=i}^{M} b'_i(j) b_2(i) && \text{(35b)} \\ &\geq \sum_{i=r}^{M} b_2(i), && \text{(35c)} \end{aligned}$$

where (35a) and (35b) come from switching the indexes and in (35c) we substitute $\sum_{j=i}^{M} b'_i(j) = 1$, because b'_i is a belief vector. For the proof of the other direction, in a similar way, we can show that if $b_1 \geq_s b_2$, there are valid belief vectors $b'_i, i \in \{1, ..., M\}$ such that (34) holds.

Let recall the definition of value function.

$$V_t(b_k) = \max_{\pi_{t:T}} V_t(b_k; \pi_{t:T}), k = 1, 2 \tag{36}$$

where $\pi_{t:T} = [\pi_t, \pi_{t+1}, ..., \pi T]$ is the sequence of policies such that $\pi_{t'}$ is the selected action at each time step $t' = t, ..., T$. Then $V_t(b_k) \geq V_t(b_k; \pi_{t:T})$ for any policy sequence $\pi_{t:T}$.

Now we only need to prove that the remaining expected reward achieved by policy sequence $\pi_{t:T}$, for belief vector b_1 is higher than b_2, *i.e.*,

$$V_t(b_1; \pi_{t:T}) \geq V_t(b_2; \pi_{t:T}) \tag{37}$$

Let's name the sequence of the belief vectors with initial vector of b_1 by $b_{1,\tau}, t \leq \tau \leq T - 1$, and the sequence of the belief vectors with initial vector of b_2 by $b_{2,\tau}$.

Considering all possible sample paths of $[B_t, B_{t+1}, ..., B_T]$, defined as $B_{t:T}$, the total remaining expected rewards for $k = 1, 2$ will be as follows:

$$V_t(b_k; \pi_{t:T}) = E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B_\tau; \pi_\tau) | b_k, \pi_{t:T}]. \tag{38}$$

Therefore, we can write $V_t(b_k; \pi_{t:T})$ as follows:

$$\begin{aligned} V_t(b_k; \pi_{t:T}) = \sum_{i=1}^{M} \{ & b_k(i) \times \\ & E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B_\tau; \pi_\tau) | B_t = i]\} \end{aligned} \tag{39}$$

So using (39) and substituting (34) we have:

$$\begin{aligned} V_t(b_1; \pi_{t:T}) - V_t(b_2; \pi_{t:T}) &= \sum_{i=1}^{M} \{b_2(i) \times \\ [\sum_{j=i}^{M} b'_i(j) E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} & R(B_\tau; \pi_\tau) | B_t = j] \\ - E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} & R(B_\tau; \pi_\tau) | B_t = i]]\} \\ = \sum_{i=1}^{M} b_2(i) [\sum_{j=i}^{M} & b'_i(j) \Delta E_{j,i}], \end{aligned} \tag{40}$$

where,

$$\begin{aligned} \Delta E_{j,i} = E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B_\tau; \pi_\tau) | B_t = j] \\ - E_{B_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B_\tau; \pi_\tau) | B_t = i]]. \end{aligned} \tag{41}$$

To prove (37), we only need to show $\Delta E_{j,i} \geq 0$. Without loss of generality, we can assume $i = j - 1 = l$. The result will be easily extendable to $j > i + 1$, by using $\Delta E_{j,i} = \sum_{l=i}^{j-1} \Delta E_{l+1,l}$.

For any sample path $B_{t:T}$ starting from l (call it $B^l_{t:T}$), there is a corresponding sample path $B^{l+1}_{t:T}$ starting from $l + 1$, such that, $B^{l+1}_\tau = B^l_\tau + 1$. So we can have the expectation over only all possible $B^l_{t:T}$ and write (41) as follows:

$$\begin{aligned} E_{B^{l+1}_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B^{l+1}_\tau; \pi_\tau)] - E_{B^l_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} R(B^l_\tau; \pi_\tau)] \\ = E_{B^l_{t:T}}[\sum_{\tau=t}^{T} \beta^{\tau-t} \{R(B^l_\tau + 1; \pi_\tau) - R(B^l_\tau; \pi_\tau)\}] \end{aligned} \tag{42}$$

Now if we show $\sum_{\tau=t}^{T} \beta^{\tau-t} \{R(B^l_\tau + 1; \pi_\tau)] - R(B^l_\tau; \pi_\tau)\} \geq 0$ for any sample path $B^l_{t:T}$, (42), therefore (40) are greater than or equal zero and (37) holds. We could use induction to prove this inequality. For the base case at horizon T, it holds using the fact that (3) is monotonically increasing with respect to B_t. Now assuming it is true for time steps $t + 1, ..., T$, we need to show it for time step t.

Now let's define the fist time that the selected action will exceed the B^l_τ, with μ, *i.e.*,

$$\mu = \min\{t \leq \tau \leq T : \pi_\tau > B^l_\tau\} \tag{43}$$

There are three different possible cases to happen:

Case I: $\pi_\tau \le B^l_\tau,\ \forall \tau \le T$

Case II: $\pi_\mu > B^l_\mu + 1 = B^{l+1}_\mu$

Case III: $\pi_\mu = B^l_\mu + 1$

Let's consider Case I, first. In this case, the actions are always lower than B^l_τ. Therefore by (3), both rewards for $B^l_{t:T}$ and $B^l_{t:T} + 1$ will be equal to the selected action. So,

$$\sum_{\tau=t}^{T} \beta^{\tau-t}\{R(B^l_\tau + 1; \pi_\tau) - R(B^l_\tau; \pi_\tau)\} = \sum_{\tau=t}^{T} \beta^{\tau-t}\pi_\tau \ge 0 \tag{44}$$

In Case II, the rewards accumulated before time μ are equal for both beliefs. Therefore,

$$\sum_{\tau=t}^{\mu-1} \beta^{\tau-t}\{R(B^l_\tau + 1; \pi_\tau) - R(B^l_\tau; \pi_\tau)\} = 0 \tag{45}$$

The rewards earned at time μ, by (3) is equal to:

$$R(B^l_\mu + 1; \pi_\mu) - R(B^l_\mu; \pi_\mu) = 1 + C \tag{46}$$

The expected value of (46) is also equal to $1+C$. After time μ, both beliefs will be restarted to $I_{B^l_\mu}$ and $I_{B^l_\mu+1}$, and from the correctness of (42) at time step μ as the induction assumption, the remaining expected reward related to B^{l+1} will be higher than the one related to B^l.

$$\begin{aligned} &E_{B^l_{\mu+1:T}}[\sum_{\tau=\mu+1}^{T} \beta^{\tau-t}\{R(B^l_\tau + 1; \pi_\tau) - R(B^l_\tau; \pi_\tau)\}] \\ &= \beta^{\mu+1-t}[V_{\mu+1}(I_{B^l_\tau+1}P) - V_{\mu+1}(I_{B^l_\tau}P)] \ge 0 \end{aligned} \tag{47}$$

Therefore, (42) which is the summation of three terms in (45), (46) multiplied by $\beta^{\mu-t}$, and (47), will be greater than or equal to zero.

For Case III, at μ the action will exceed B^l_μ and the belief will be restarted to $I_{B^l_\mu}P$, but the other sequence's belief will change as $b_{\mu+1} = T_{\pi_\mu} b_\mu P$. The difference between rewards accumulated before time μ and at time μ are equal to 0 and $C + 1$, respectively, similar to (45) and (46). Now for times after μ we have:

$$\begin{aligned} &E_{B^l_{\mu+1:T}}[\sum_{\tau=\mu+1}^{T} \beta^{\tau-t}\{R(B^l_\tau + 1; \pi_\tau) - R(B^l_\tau; \pi_\tau)\}] \\ &= \beta^{\mu+1-t}[V_{\mu+1}(T_{\pi_\mu} b_\mu P) - V_{\mu+1}(I_{B^l_\tau}P)] \ge 0 \end{aligned} \tag{48}$$

where the last inequality comes from the correctness of (37) at time $\mu + 1$ as the assumption of the induction, considering the new belief $T_{\pi_\mu} b_\mu$ and $I_{B^l_\tau}P$. Note B^l_μ is the lowest index of non-zero beliefs in $T_{\pi_\mu} b_\mu$. Therefore, (42) is greater than or equal to zero, so are (41) and (40), and thus (37) is true for time t and the proof is complete.

Now assume $\pi^*_{t:T} = \arg\max_{\pi_{t:T}} V_t(b_2; \pi_{t:T})$ is the optimal policy sequence for the initial belief vector b_2 at time t. We have:

$$\begin{aligned} V_t(b_1) &\ge V_t(b_1; \pi^*_{t:T}) \\ &\ge V_t(b_2; \pi^*_{t:T}) = V_t(b_2), \end{aligned} \tag{49}$$

and this completes the proof of Lemma 4. ■

Proof of Lemma 5:

To prove the shifting property of immediate expected reward we have:

$$\bar{R}(b^\alpha; r) = r\sum_{i=r}^{M} b^\alpha(i) + \sum_{i=1}^{r-1} b^\alpha(i)[(1 + C)i - Cr] \tag{50a}$$

$$\begin{aligned} &= (r' + \alpha)\sum_{i'=r'}^{M} b(i') \\ &+ \sum_{i'=1}^{r'-1} b(i')[(1 + C)(i' + \alpha) - Cr'] \end{aligned} \tag{50b}$$

$$= \bar{R}(b; r') + \alpha[\sum_{i'=r'}^{M} b(i') + \sum_{i'=1}^{r'-1} b(i')] = \bar{R}(b; r') + \alpha, \tag{50c}$$

where, (50b) is achieved by substituting i and r with $i'+\alpha$ and $r' + \alpha$, respectively. Also we already know that $b^\alpha(i' + \alpha) = b(i')$. For the myopic actions, using (50c) we have:

$$\begin{aligned} &r^{myopic}(b^\alpha) = \arg\max_r \bar{R}(b^\alpha; r) \\ &= \arg\max_r\{\bar{R}(b; r - \alpha) + \alpha\} = r^{myopic}(b) + \alpha. \end{aligned} \tag{51}$$

■

Proof of Lemma 6:

To prove (20), we use induction. For horizon T, it's similar to (18) with $t = T$. Now assuming it's true for $t + 1$, we will consider time step t:

$$\begin{aligned} V_t(b^\alpha; r) &= \bar{R}(b^\alpha; r) + \beta[\sum_{i=r}^{M} b^\alpha(i) V_{t+1}(T_r b^\alpha P) \\ &+ \sum_{i=1}^{r-1} b^\alpha(i) V_{t+1}(I_i P)] \\ &= \bar{R}(b; r) + \alpha + \beta[\sum_{i'=r-\alpha}^{M} b(i') V_{t+1}(T_r b^\alpha P) \\ &+ \sum_{i'=1}^{r-\alpha-1} b(i') V_{t+1}(I_{i'+\alpha}P)] \end{aligned} \tag{52a}$$

In (52a), we use (18) and the shifting property of belief vectors and substitute $i - \alpha$ with i'. Using (20) at $t + 1$ we have:

$$\begin{aligned} V_t(b^\alpha; r) &= \bar{R}(b; r) + \alpha \\ &+ \beta\sum_{i'=r-\alpha}^{M} b(i')[V_{t+1}(T_{r-\alpha} bP) + \alpha\frac{1 - \beta^{T-t-1}}{1 - \beta}] \\ &+ \beta\sum_{i'=1}^{r-\alpha-1} b(i')[V_{t+1}(I_{i'}P) + \alpha\frac{1 - \beta^{T-t-1}}{1 - \beta}] \end{aligned} \tag{53a}$$

$$\begin{aligned} &= V_t(b; r - \alpha) + \alpha + \alpha\beta\sum_{i=1'}^{M} b(i')\frac{1 - \beta^{T-t-1}}{1 - \beta} \\ &= V_t(b; r - \alpha) + \alpha\frac{1 - \beta^{T-t}}{1 - \beta} \end{aligned} \tag{53b}$$

To prove (21), using (20), we have:

$$\begin{aligned} r_t^{opt}(b^{\alpha}) &= \arg\max_r V_t(b^{\alpha}; r) \\ &= \arg\max_r \{V_t(b; r-\alpha) + \alpha \frac{1-\beta^{T-t}}{1-\beta}\} \\ &= \arg\max_r V_t(b; r-\alpha) \\ &= \arg\max_r V_t(b; r) + \alpha = r_t^{opt}(b) + \alpha. \end{aligned} \tag{54}$$

Using (20), (22) and maximizing over actions, we obtain (22). ■

Proof of Lemma 7:

Consider a belief vector b^u in the unlimited state set and its correspond b^l in the limited state set $\mathcal{M}$. If b^u has non-zero elements above M, *i.e.*, the upper edge of $\mathcal{M}$, $b^l(M) = \sum_{i \geq M} b^u(i)$, and $b^l(j) = b^u(j), j < M$. Therefore facing the upper edge effect results in $b^l \leq_s b^u$. Similarly, is b^u has non-zero elements below 1, *i.e.*, the lower edge of $\mathcal{M}$, $b^l(1) = \sum_{i \leq 1} b^u(i)$, and $b^l(j) = b^u(j), j > 1$. Thus facing with the lower edge effect results in $b^l \geq_s b^u$.

Now using Lemma 6, if the belief vector b_t and its shifted version $b_t^{\alpha}, \alpha > 0$ are in the unlimited state set, between their value functions the equation (22) holds. Lets b_{τ} and b_{τ}^{α} denote the sequence of belief vectors constructed by updating b_t and b_t^{α}, respectively, according to their optimal actions at all time steps $\tau = t+1, ..., T$. Then $b_{\tau}^{\alpha}, \tau = t+1, ..., T$ are shifted version of b_{τ}. Now with limiting the state set,there are some possibilities. If none of the $b_{\tau}, b_{\tau}^{\alpha}, \tau = t+1, ..., T$ face with the edge effects and (22) will still hold. Otherwise, b_{τ}^{α} will face with the upper edge effect sooner than b_{τ}, or b_{τ} will face with the lower edge effect sooner than b_{τ}^{α}. In both cases, combining the above argument about the FOSD orderings, the results of Lemma 4 and (22), we can conclude the inequality (23). ■

APPENDIX B

Proof of Theorem 1:

The immediate expected reward and the future expected reward achieved by the action $r < r^{myopic}$ are less than those achieved by r^{myopic}, as results of Lemma 1 and Lemma 3, respectively. Combining them as (5b), we get $V_t(b; r) \leq V_t(b; r^{myopic})$ which means r cannot be optimal and thus $r^{opt} \geq r^{myopic}$. ■

Proof of Theorem 2:

To achieve the upper bound on the optimal action we will show that:

$$\Delta V_t(b; r) = \Delta R(b; r) + \beta \Delta V_t^f(b; r) \leq 0, \ \forall r \geq r^{ub}. \tag{55}$$

By recalling (24) we know

$$\Delta R(b; r) = 1 - (1 + C) \sum_{i=1}^{r} b(i) \leq 0, \quad \forall r \geq r^{myopic}. \tag{56}$$

Moreover, using (32),

$$\begin{aligned} \Delta V_t^f(b; r) = & \sum_{i=r+1}^{M} b(i)[V_{t+1}(T_{r+1}bP) - V_{t+1}(T_r bP)] \\ & + b(r)[V_{t+1}(I_r P) - V_{t+1}(T_r bP)], \end{aligned} \tag{57}$$

and according to lemma 3 we have $\Delta V_t^f(b; r) \geq 0$ for all r. Hence (55) is equivalent to:

$$\beta \Delta V_t^f(b; r) \leq |\Delta R(b; r)|, \quad \forall r \geq r^{ub}. \tag{58}$$

To this end, it is sufficient to find an upper bound for $\beta \Delta V_t^f(b; r)$ such that it is less than $|\Delta R(b; r)|$ for all r greater than a specific action, *i.e.*, our desired upper bound, r^{ub}. Applying Lemma 4 one $bp \geq_s I_{r^l} P$, we have:

$$V_t(bP) \geq V_t(I_{r^l} P), \tag{59}$$

Now by above inequality and the fact that r is the lowest non-zero index of the belief vector $T_r b$, the second term in (57) is less than zero and we can bound $\Delta V_t^f(b; r)$ as follows:

$$\Delta V_t^f(b; r) \leq \sum_{i=r+1}^{M} b(i)[V_{t+1}(T_{r+1}bP) - V_{t+1}(T_r bP)]. \tag{60}$$

From the convexity of value function proved in lemma 2 and (59) we get:

$$\begin{aligned} & V_{t+1}(T_{r+1}bP) - V_{t+1}(T_r bP) \\ & \leq \sum_{i=r+1}^{r^h} b'(i) V_{t+1}(I_i P) - V_{t+1}(I_r P) \\ & \leq \sum_{i=r+1}^{r^h} \{b'(i)[V_{t+1}(I_i P) - V_{t+1}(I_r P)]\}, \end{aligned} \tag{61}$$

where $b' = T_{r+1}b$. Now from (23) and substituting $b = I_i P$ and $\alpha = j - i$, it follows that:

$$\begin{aligned} & V_{t+1}(T_{r+1}bP) - V_{t+1}(T_r bP) \\ & \leq \sum_{i=r+1}^{r^h} \{b'(i) \frac{(i-r)(1-\beta^{T-1})}{1-\beta}\} \\ & = \frac{1-\beta^{T-1}}{1-\beta} [\sum_{i=r+1}^{r^h} i b'(i) - r] \\ & = \frac{1-\beta^{T-1}}{1-\beta} [\bar{b}' - r], \end{aligned} \tag{62}$$

where $\bar{b}' = \sum_{i=r+1}^{r^h} i b'(i) = \frac{\sum_{i=r+1}^{r^h} i b(i)}{\sum_{i=r+1}^{r^h} b(i)} = \frac{\bar{S}_r}{S_r}$.

Accordingly, we may bound (60) using (62) as follows:

$$\Delta V_t^f(b; r) \leq S_r \frac{1-\beta^{T-1}}{1-\beta} [\frac{\bar{S}_r}{S_r} - r]. \tag{63}$$

To obtain (58) using (63) it suffices to make the following inequality satisfied:

$$\beta S_r \frac{1-\beta^{T-1}}{1-\beta}[\frac{\bar{S}_r}{S_r} - r] \leq (1+C)\sum_{i=1}^{r} b(i) - 1$$
$$= C - (1+C)S_r,\ \forall r \geq r^{ub}. \quad (64)$$

Applying the definition of $f(\beta)$ in (14) and doing some straightforward manipulations, we can get:

$$f(\beta)\bar{S}_r + [(1+C) - f(\beta)r]S_r - C \leq 0,\ \forall r \geq r^{ub}. \quad (65)$$

r^{ub} is the smallest r satisfying (65), as mentioned in (13), and the proof is complete.

■