Proceedings of the International Conference on Industrial Engineering and Operations Management
Paris, France, July 26-27, 2018

Percentile Policies for Inventory Problems with Partially Observed Markovian Demands

Farzaneh Mansourifard
Department of Engineering & Technology
Alzahra University
Tehran, Iran
farzaneh.mansourifard@gmail.com

Parisa Mansourifard, Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, California, 90089 USA
parisama@usc.edu, bkrishna@usc.edu

Abstract

We study a set of inventory control problems with correlated demands over different time periods. On the other hands, we relax the assumption of fully observation of the demand at the end of each time period. In other words, we consider the case of partially observed (censored) demand in the context of a multi-period inventory problem. If the demand in a period is larger than the inventory level, we don't observe the unmet demand. Otherwise, the demand is fully observed and the leftover inventory is carried over to the next period. Formulating the problem as a Partially Observable Markov Decision Process provides a dynamic program (DP) to minimize the total expected cost. Unfortunately, the corresponding DP is defined on an uncountable state space, with little hope for a computationally feasible solution. We present an interesting heuristic policy with a percentile threshold structure which outperforms the myopic policy and performs close to the optimal policy. We derive its performance guarantee and evaluate it using numerical simulations.

Keywords

Multi-period inventory management, Markovian demand, censored demand, Dynamic Programming

1. Introduction

Inventory control is one of the important topics in operations research and management and it has been studied by many researchers (Qin, 2011). In this kind of problems, the demand for some good is assumed to follow a stochastic process and at the beginning of each decision epoch the decision-maker decides on the inventory level (i.e. how many items to store) in order to satisfy the demand. As one of the challenging problems in inventory control, many studies have been focused on different distributions of the demand. Most inventory models in these studies assume that demands are independent and identically distributed over different time periods (e.g. Besbes 2010). However, in recent years, it has been observed that this assumption might not hold in practice (Tai 2016), thus, there have been research papers on the inventory problems with correlated demands over time (Hu 2016, Alwan 2016). For instance, in some studies the demand is assumed to be Markov-modulated (Hu, 2016), or Autoregressive (Alwan 2016).

Another challenging problem in this field is about the observability of the demand. In other words, if the demand is higher than the sale, the unmet demand might not be fully observable. In some inventory systems such as retail stores, unmet demand of inventory is lost and cannot be observed or recorded. In literature, this problem is referred as censoring or partial observability (e.g. Lu 2008, Bisi 2011).

In this paper, we address the two aforementioned challenging problems. We study the inventory control problem when the demand is partial observable, and correlated over time periods as a Markovian process. Therefore, we face with a Partial Observable Markov Decision Process (POMDP) problem. As such the solution of this problem can be

characterized via a dynamic programming (DP), however, it is computationally complex and a POMDP is generally known to be P-SPACE hard (Papadimitriou 1987). In literature, e.g. Bensoussan 2007 and 2008, a similar problem set is studied and the "existence" of an optimal policy is shown. In our previous work, we proposed a sub-optimal solution for a POMDP to solve the perishable inventory control problem (Mansourifard 2017). In particular, we introduced a new class of heuristic percentile policies with percentile threshold structure, and evaluated their performance. In this paper, we consider an extension to a multi-period inventory control with censored Markovian demand in which the leftover inventory is carried over to the next period. We present the heuristic policy for this problem and evaluate its performance using the lower bound derived on the cost of the optimal policy.

The remainder of the paper is organized as follows: In section 2, we review the related works. The problem formulation is given in section 3 followed by the dynamic programming formulation in section 4. Section 5 presents the heuristic policy and its performance bound. The simulation results are given in section 6. Finally, we conclude the paper in section 7.

2. Related Literature

Most of the inventory control literature (e.g. Ding 2002, and Bensoussan 2009) assume that the demand process is independent and identical distributed (i.i.d) at different time periods. Some prior works (such as Negoescu 2008, Besbes 2013) consider the case where the demand distribution is i.i.d but unknown, so the learning plays an important role in estimating the distribution and making the decision. For instance, in Besbes 2013, the demand distribution is estimated from historical data. They show that the optimal policy has a percentile structure and characterize the implications of partial observations on the performance of the optimal policy in both discrete and continuous settings. However, in recent years, it has been observed that the demand distribution is not necessarily i.i.d and it can have correlation over time (Tai 2016). For example, Hu 2016 studied the inventory control problem with Markov-modulated demand. Note that in most of these papers, the demand is assumed to be fully observed.

In some other literature works such as Lu 2008, Chen 2010, and Bisi 2011, the inventory problem with partially observed (censored) i.i.d. demand has been studied. In Bisi 2011, a Bayesian scheme is employed to dynamically update the demand distribution for the problem with storable or perishable inventory. They show that the Weibull is the only newsvendor distribution for which the optimal solution can be expressed in scalable form. In Lu 2008, the perishable inventory control problem with censored demand is studied in which the demand distribution is assumed to be i.i.d. but unknown. They use Bayesian approach to update the distribution parameters periodically based on the censored historical sales data. Chen 2010 studied the non-perishable inventory control problem with censored and i.i.d. demand. They developed bounds and heuristics for such a problem.

Furthermore, there are some research papers which study the inventory control problem with censored and temporally correlated demands. In Bensoussan 2007, a perishable inventory management problem with memory (Markovian) demand process is considered. In their work, some structural properties of the optimal actions relative to the myopically optimal actions are obtained. And in Bensoussan 2008, they extended the work to the non-perishable inventory. In these papers, the existence of an optimal policy is shown. In their work, some structural properties of the optimal actions relative to the myopically optimal actions are obtained. In this paper, in contrast, we focus on the design and analysis of a class of heuristic policies. In particular, we present the class of percentile policies and evaluate their performance. In addition, we present a lower bound on the cost of the optimal policy which can be computed with low complexity and give a measure for how close our heuristic policies are to the optimal policy.

In our previous works (Mansourifard 2013, and 2015), we studied another version of this problem with no carry-over in a network congestion control context and derived lower and upper bounds on the optimal policy. In this work, we study the inventory control problem in which the demand is censored and Markovian and the left-over inventory is carried over to the next time period.

3. Problem Formulation

In this problem, the demand is a Markovian process which is only partially observable to the decision-maker and the action that the decision-maker must take is the quantity to be ordered to increase the inventory level. We consider a discrete-time finite-state Markov process whose state, denoted by d_t, is the demand amounts evolving based on a known transition matrix over a finite horizon, T. At each time step (period) t, the decision-maker selects an ordered quantity based on the history of observations and pays a cost which is a function of the total inventory level (i.e. previous inventory level plus ordered quantity) and the actual demand B_t. If the total inventory level is higher than

the actual demand, the demand can be fully observed based on the number of sold items, otherwise, only the fact that the demand was higher than the inventory level will be revealed (partial observation).

The objective is to select the sequential actions (policy) such that the total expected cost accumulated over the horizon is minimized. Selecting an ordered quantity which makes the inventory level higher than the actual demand causes the over-utilization cost, but gives full information about the actual demand. On the other hand, selecting an ordered quantity which does not increase the inventory level higher than the actual demand causes under-utilization cost, and only gives partial information about the actual demand. Therefore, the decision-maker faces with a trade-off between selecting the ordered quantity which minimizes the immediate expected cost and selecting higher ordered quantity to earn more information to minimize the future expected cost.

Since we do not get full observation all the time, we formulate our problem within a POMDP-based framework defined as follows:

- State: The state of Markov process, d_t, is one of the elements of a finite state set denoted by $\mathcal{M} = \{0,1,\dots,M\} \subset \mathbb{Z}$.
- State transition: The transition probabilities of the actual demand d_t over time are assumed to be known and stationary and indicated by a transition probability matrix, P. This is an $|\mathcal{M}| \times |\mathcal{M}|$ matrix with elements $P_{i,j} = Pr(d_{t+1} = j | d_t = i), i, j \in \mathcal{M}, \forall t$ which indicates the probability of moving from state i at a time step to the actual demand j at the next time step.
- Action: At each time step, we choose an action $q_t \in \mathcal{M}$ as the ordered quantity. Note that the set of actions are equal to the set of demands. We have an inventory level, denoted by L_t, which is the leftover inventory from previous time steps.
- Observed information: The observed information at time step t is defined by the event $o_t(q_t + L_t) \in \mathrm{O}$ which is a function of the inventory level, ordered quantity and the actual demand. The possible events corresponding to the action q_t is as follows:

- $o_t(q_t + L_t) = \{d_t = i\}, i \in \{0,\dots,L_t + q_t - 1\}$ is the event of fully observing d_t. This corresponds to the ordering of the quantity which increases the inventory level higher than d_t.

- $o_t(q_t + L_t) = \{d_t \geq q_t + L_t\}$ is the event of partial observing that d_t is larger than or equal to the inventory level plus ordered quantity.

- Cost: The immediate cost paid at time step t is a mapping $\mathrm{C}: \mathcal{M} \times \mathcal{M} \times \mathrm{O} \to \mathbb{R}$, and depends on the inventory level L_t, the demand d_t, and the ordered quantity q_t. Therefore, the immediate cost function is given by:

$$C(d_t, L_t; q_t) = c_0 q_t + \begin{cases} c_u(L_t + q_t - d_t) & if \quad d_t \leq L_t + q_t \\ c_l(d_t - L_t - q_t) & if \quad d_t \geq L_t + q_t \end{cases} \qquad (1)$$

where c_u and c_l are the over-utilization (holding) and the under-utilization (shortage) cost per unit, respectively, and c_o is the ordering cost per unit.

4. Dynamic Programming Formulation

We represent the decision problem based on the decision-maker's belief, i.e. his posterior probability conditioned on past actions and observations. In other words, we define the state to be the belief vector representing the conditioned probability distribution on the hidden demand d_t at each time step and minimize the expected cost-to-go corresponding to the belief. Let the conditioned probability distribution of the demand (assuming a finite state set), given all past observations, is denoted by a belief vector $b_t = [b_t(0),\dots,b_t(M)]$, with elements of $b_t(k) = \Pr|past\ observaitons),\ k \in \mathcal{M}$. In other words, b_t represents the probability distribution of d_t over all possible demands of $\mathcal{M}$. The set of all possible belief vectors is denoted by D.

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-maker may only make the decision based on the belief vector. It can be shown that the belief vector is a sufficient statistic of the complete observation history.

The belief updating $\mathcal{M} \times \mathrm{O} \times D \to D$ maps current belief vector, updated inventory level, and the observation to the belief vector for the next time step:

$$b_{t+1} = \begin{cases} T_{L_t+q_t}[b_t]P & if \quad o_t = \{d_t \geq L_t + q_t\}, \\ I_i P & if \quad o_t = \{d_t = i\}, \end{cases} \qquad (2)$$

where I_i is the $M+1$-dimensional unit vector with 1 in the i-th position and 0 otherwise. Note that $I_i P$ is equivalent to the i-th row of matrix P, i.e. $P_{i,.}$. T_a is a non-linear operation on a belief vector b, as follows:

$$T_a[b](i) = \begin{cases} 0 & if\ i < a, \\ \dfrac{b(i)}{\sum_{j=a}^{M} b(j)} & if\ i \geq a. \end{cases} \qquad (3)$$

The inventory level will be updated as follows:

$$L_{t+1} = \begin{cases} 0 & if\ o_t = \{d_t \geq L_t + q_t\}, \\ L_t + q_t - i & if\ o_t = \{d_t = i\}, \end{cases} \qquad (4)$$

Figure 1 shows the POMDP models for this problem.

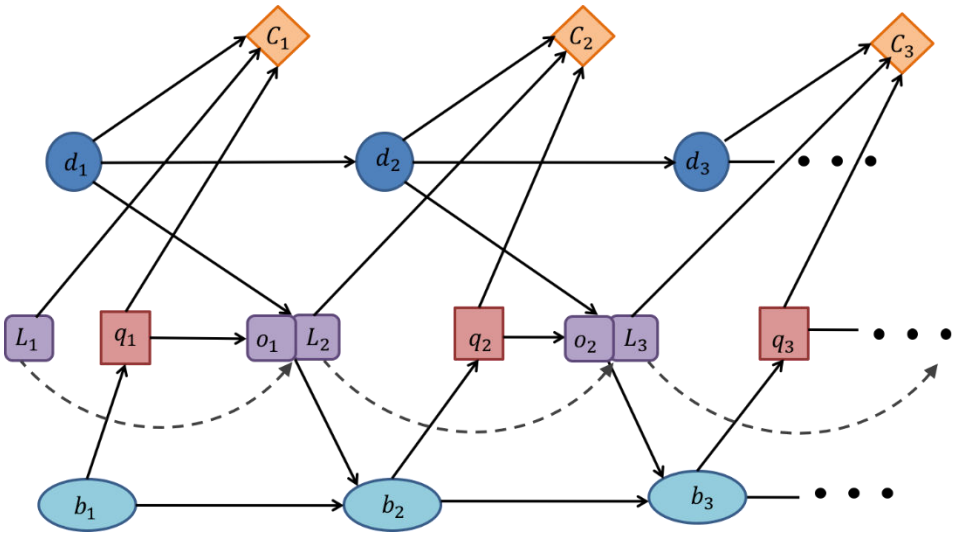

Fig. 1. The POMDP model

The immediate expected cost, caused by selecting the ordered quantity q_t and based on the belief vector b_t and the leftover inventory L_t is obtained by taking expectation of (1), as follows:

$$\bar{C}(b_t, L_t; q_t) = \sum_{i \in \mathcal{M}} b_t(i) C(i, L_t; q_t)$$

$$= c_0 q_t + c_l \sum_{i=L_t+q_t}^{M} b_t(i)(i - L_t - q_t) + c_u \sum_{i=0}^{L_t+q_t-1} b_t(i)(L_t + q_t - i). \qquad (5)$$

The goal is to minimize the total expected cost in the horizon T, over all admissible policies π, given by

$$\min_{\pi} J_T^{\pi}(b_1, L_1) = \min_{\pi} \mathbb{E}\{\sum_{t=1}^{T} C(d_t, L_t; q_t) | b_1\}, \qquad (6)$$

where b_1 and L_1 are the initial belief vector and the initial inventory level, respectively. $J_T^{\pi}((b_1, L_1)$ is the total expected cost accumulated over the horizon T under policy π. The policy π specifies a sequence of functions $\pi_1, \ldots, \pi_T$,where $\pi_t$is the decision rule and maps a belief vector b_t and inventory level L_t to an ordered quantity at time step t, i.e., $\pi_t: D \times \mathcal{M} \rightarrow \mathcal{M}$, $\mathcal{M} \quad q_t = \pi_t(b_t, L_t)$. The optimal policy denoted by π^{opt} is a policy which minimizes (6) 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 hold:

$$V_t(b_t, L_t) \coloneqq \min_{q_t} V_t(b_t, L_t; q_t), \tag{7a}$$
$$V_T(b_T, L_T; q_T) = \bar{C}_T(b_T, L_T; q_T), \tag{7b}$$
$$V_t(b_t, L_t; q_t) \coloneqq \bar{C}(b_t, L_t; q_t) + \mathbb{E}\{V_{t+1}(b_{t+1}, L_{t+1}) | L_t, q_t, b_t\}, \quad t < T \tag{7c}$$

where b_{t+1} and L_{t+1} are the updated belief vector and inventory level, respectively. They can be computed given the ordered quantity q_t and observation o_t as shown in (2) and (4). The value function $V_t(b_t, L_t)$ is the minimum expected cost-to-go when the current belief vector is b_t and the inventory level is L_t. Note that $V_t(b_t, L_t; q)$ is the expected cost-to-go after time t under belief b_t, inventory level L_t and the ordered quantity q at time t and following the optimal policy for time t+1 onward, with updated belief vector and inventory level according to the ordered quantity q. The future expected cost can be computed as follows:

$$\begin{aligned} &\mathbb{E}\{V_{t+1}(b_{t+1}, L_{t+1}) | L_t, q_t, b_t\} \\ &= \sum_{i=L_t+q_t}^{M} b_t(i) V_{t+1}\big(T_{L_t+q_t}[b_t]P, 0\big) \\ &+ \sum_{i=0}^{L_t+q_t-1} b_t(i) V_{t+1}\big(P_{i,.}, L_t + q_t - i\big). \end{aligned} \tag{8}$$

Note that for all $t = 1, \dots, T$, $V_t(b_t, L_t) = min_\pi J_{T-t}^\pi(b_t, L_t)$ with probability 1. In particular, $V_1(b_1, L_1) = J_T^\pi(b_1, L_1)$ A policy π^{opt} is optimal if for $t = 1, \dots, T$; $r_t^{opt}(b_t, L_t)$ achieves the minimum in (7a), denoted by:

$$q_t^{opt}(b_t, L_t) \coloneqq \arg\min_{q \in \mathcal{M}} V_t(b_t, L_t; q). \tag{9}$$

5. Heuristic Policy and its Performance Bound

5.1 Percentile Threshold Policies

Since finding the optimal policy is computationally intractable for large horizons, we consider specific form of heuristic policies which have percentile threshold structure as follows:

$$q^{Percentile}(b_t, L_t) = \min\{q \in \mathcal{M} : \sum_{i=0}^{L_t+q} b_t(i) \geq h^{Percentile}(b_1, L_1)\}, \tag{10}$$

where the threshold $h^{Percentile}(b_1, L_1)$ is a function of the initial belief vector b_1 and the initial inventory level L_1. From now on we will call this form of heuristic policies, percentile policies. The reason to consider this specific form of policies is that later in this paper we derive a lower and an upper bound on the optimal policy (with some condition on the parameters) which both have percentile threshold structures and conjecture that there may be a good approximation for the optimal policy with the same structure.

5.2 Performance Bound of PT Policies

In this section, we present a performance guarantee for percentile policies in the following theorem. This performance guarantee is used to evaluate the heuristic percentile policies in the Simulation section.

Theorem 1. The performance bound on the percentile policy with threshold $h^{Percentile}$ is given by:

$$\frac{J_{T-1}^{Percentile}(b_1, L_1)}{J_{T-1}^{opt}(b_1, L_1)} \leq \frac{J_{T-1}^{Percentile}(b_1, L_1)}{J_{T-1}^{FO}(b_1, L_1)} = \frac{V_1^{Percentile}(b_1, L_1)}{V_1^{FO}(b_1, L_1)}, \tag{11}$$

$$V_t^{FO}(b_t, L_t) = \min_q \bar{C}(b_t, L_t; q) + \sum_{i=0}^{M} b_t(i) V_{t+1}^{FO}\big(P_{i,.}, (L_t + q - i)^+\big),$$

$$V_t^{Percentile}(b_t, L_t) = \Gamma_{T-t}(b_t, L_t, h^{Percentile})$$
$$+ \sum_{i=0}^{L_t+q^{Percentile}(b_t,L_t)-1} b_t(i) V_{t+1}^{Percentile}\left(P_{i,.}, L_t + q^{Percentile}(b_t, L_t) - i\right)$$
$$+ \sum_{\tau=t+1}^{T-1} [A_{t,\tau} \sum_{i=0}^{q^{Percentile}(b_\tau,0)-1} b_\tau(i) V_{\tau+1}^{Percentile}(P_{i,.}, q^{Percentile}(b_\tau, 0) - i)] \qquad (12)$$

Where $b_\tau = T_{L_{\tau-1}+q_{\tau-1}}[b_{\tau-1}]$ and,

$$\Gamma_{T-t}(b_t, L_t, h^{Percentile}) \coloneqq \bar{C}\left(b_t, L_t; q^{Percentile}(b_t, L_t)\right) +$$
$$+ \sum_{\tau=t+1}^{T} A_{t,\tau} \bar{C}\left(b_\tau, 0; q^{Percentile}(b_\tau, 0)\right),$$
$$A_{t,\tau} \coloneqq \sum_{i=L_t+q_t}^{M} b_t(i) \prod_{t'=t+1}^{\tau-1} [\sum_{i=q_{t'}}^{M} b_{t'}(i)], \qquad (13)$$

such that $\bar{C}(b, L; q)$is the expected immediate cost, defined in (5).

Note that $V_{\tau+1}^{PT}(P_{i,.})$and $V_{t+1}^{FO}(P_{i,.})$ can be computed recursively from (12). To proof the above theorem, we need the following proposition.

Proposition 1. The cost-to-go of the optimal policy is lower bounded by the cost-to-go of the full observation (FO) case under the same belief vector, i.e.,

$$V_t(b_t, L_t) \geq V_t^{FO}(b_t, L_t). \qquad (14)$$

Note that FO scenario corresponds to simpler case where in both cases of under/over-utilization the actual demand could fully observed. In other words, there is no asymmetry in the observation. Since more information reveals at each time, the total cost could be less than the total cost of partial observation case. This is given in the following proposition. See Appendix A for proof.

5.3 Optimal Percentile Threshold Policy

In this section, we introduce the optimal percentile threshold policy, which chooses a threshold providing the minimum cost-to-go for the given initial belief vector among all possible thresholds. We will show in Simulation Section that this policy outperforms the myopic policy and performs close enough to the optimal policy. Among all percentile threshold policies, the *PT- opt* policy provides the minimum cost-to-go which is given by:

$$r^{PT-opt}(b_t, L_t) = \min\{r \in \mathcal{M}: \sum_{i=0}^{L_t+r} b_t(i) \geq h^{PT-opt}(b_1, L_1)\}, \qquad (15)$$

such that

$$h^{PT-opt}(b_1, L_1) = arg \min_{h^{PT}} J_{T-1}^{PT}(b_1, L_1), \qquad (16)$$

where $J_{T-1}^{Percentile}(b_1, L_1)$ is the total expected cost equal to $V_1^{Percentile}(b_1, L_1)$ defined in (12) achieved by selecting the actions corresponding to the threshold $h^{Percentile}(b_1, L_1)$ at all the time steps $t = 1, \dots, T$.

6. Numerical results

We present some numerical results to evaluate the performance of the introduced heuristic policy, PT-opt. The simulation parameters, except in the figures that their effect is considered, are fixed as follows: the number of states M = 9, the under-utilization cost coefficient $c_u = 0.5, c_0 = 1$, and the transition probabilities given by:

$$P = \begin{bmatrix} .6 & .1 & .2 & .1 & 0 & 0 & 0 & 0 & 0 & 0 \\ .4 & .2 & .1 & .2 & .1 & 0 & 0 & 0 & 0 & 0 \\ .3 & .1 & .2 & .1 & .2 & .1 & 0 & 0 & 0 & 0 \\ 0 & .3 & .1 & .2 & .1 & .2 & .1 & 0 & 0 & 0 \\ 0 & 0 & .3 & .1 & .2 & .1 & .2 & .1 & 0 & 0 \\ 0 & 0 & 0 & .3 & .1 & .2 & .1 & .2 & .1 & 0 \\ 0 & 0 & 0 & 0 & .3 & .1 & .2 & .1 & .2 & .1 \\ 0 & 0 & 0 & 0 & 0 & .3 & .1 & .2 & .1 & .3 \\ 0 & 0 & 0 & 0 & 0 & 0 & .3 & .1 & .2 & .4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & .1 & .6 \end{bmatrix}$$

Figure 2 and 3 show the performance bounds of percentile policies versus under-utilization cost, c_l and horizon, T, respectively. As it is shown in Figure 2, for larger c_l, the performance bound is tight and it is less than 1.5. For smaller c_l, the percentile policy outperforms the myopic policy with larger gap. On the other hand, figure 3 shows that for small horizon, T, both our heuristic policy and the myopic policy perform close to the optimal and as the horizon increase, the performance bound of our heuristic stays around 1.7 but the performance bound myopic policy increases up to 2.8.

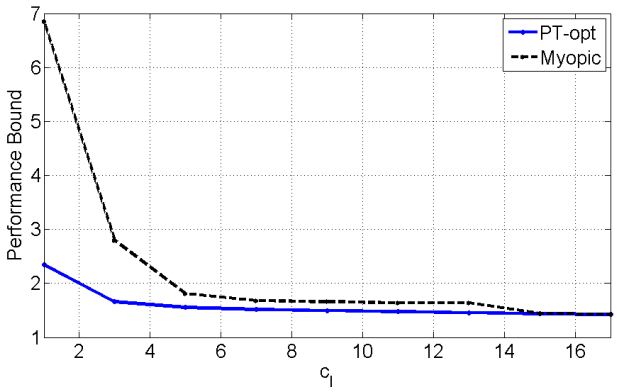

Fig. 2. The performance bound of percentile policies versus under-utilization cost, c_l, for $M = 9, c_u = 0.5, c_0 = 1, T = 20$.

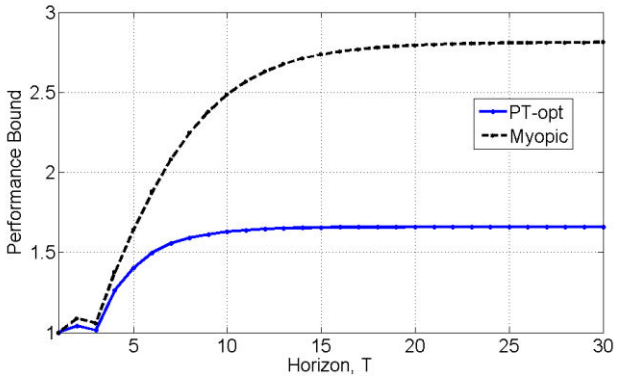

Fig. 3. The performance bound of percentile policies versus horizon T, for $M = 9, c_u = 0.5, c_0 = 1,\ c_l = 3$.

Figure 4 and 5 show the threshold of percentile policies versus under-utilization cost, c_l, and horizon, T, respectively. As figure 4 shows, our heuristic policy prefers to choose a threshold close to one, in other words it behaves aggressively to increase the chance of full observation. But the myopic policy chooses a very small threshold for small c_l, to decrease the immediate cost by acting conservatively. Furthermore, based on figure 5, our heuristic policy chooses higher thresholds for larger horizon, since it could increase the chance of getting full

observation and decreasing the future cost, but the myopic policy does not care about the future cost and chooses the same threshold for any horizon.

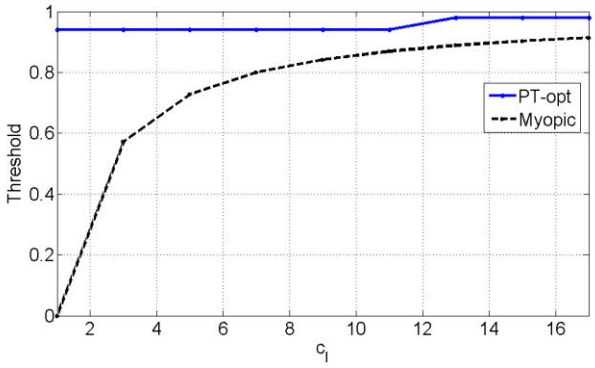

Fig. 4. The threshold of percentile policies versus under-utilization cost, c_l, for $M = 9, c_u = 0.5, c_0 = 1, T = 20$.

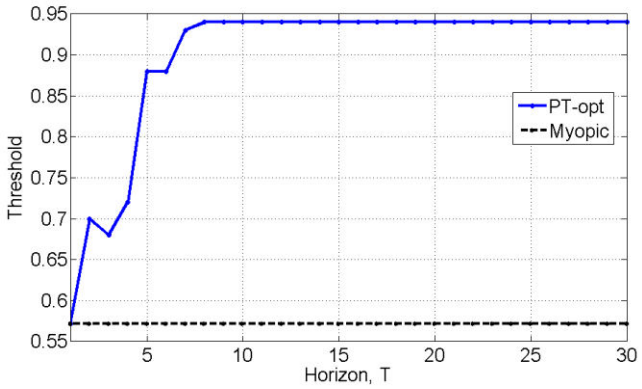

Fig. 5. The threshold of percentile policies versus horizon T, for $M = 9, c_u = 0.5, c_0 = 1,\ c_l = 3$.

7. Conclusion

We have studied a set of inventory control problems with Markovian demands over different time periods which can only be partially (censored) at the end of each period. If the demand in a period is larger than the inventory level, we don't observe the unmet demand. Otherwise, the demand is fully observed and the leftover inventory is carried over to the next period. We formulated the problem as a Partially Observable Markov Decision Process and since the corresponding DP is defined on an uncountable state space, with little hope for a computationally feasible solution, we presented an interesting heuristic policy with a percentile threshold structure which outperforms the myopic policy and performs close to the optimal policy. We derived its performance guarantee and evaluated it using numerical simulations.

As future works, we aim to identify the conditions where our heuristic policy is optimal. We can also consider a more complicated scenario where the transition matrix is unknown and needs to be learned over time. It will be interesting to study other correlation between demand overtime beside the Markovian relationship.

Acknowledgements

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". Parisa Mansourifard was supported by AAUW American Dissertation Completion Fellowship for 2015-2016.

References

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

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

Bensoussan, A., Akanyıldırım, M. C¸ Sethi, S. P., "A multiperiod newsvendor problem with partially observed demand", Mathematics of Operations Research, vol. 32, no. 2, pp. 322–344, 2007.

Negoescu, D., Frazier, P., Powell, W., "Optimal learning policies for the newsvendor problem with censored demand and unobservable lost sales", URL: http://people.orie.cornell.edu/pfrazier/pub/learning-newsvendor.pdf, 2008.

Bensoussan, A., Akanyıldırım, M. C¸ Minjarez-Sosa, J. A., Royal, A., Sethi, S. P., "Inventory problems with partially observed demands and lost sales", Journal of Optimization Theory and Applications, vol. 136, no. 3, pp. 321–340, 2008.

Lu, X., Song, J., Zhu, K., "Analysis of perishable-inventory systems with censored demand data", Operations Research, vol. 56 no. 4, pp. 1034-1038, 2008.

Bensoussan, A., Akanyıldırım, M. C¸ Sethi, S. P., "Technical note-a note on the censored newsvendor and the optimal acquisition of information", Operations Research, vol. 57, no. 3, pp. 791–794, 2009.

Chen, L., "Bounds and heuristics for optimal bayesian inventory control with unobserved lost sales", Operations research, vol. 58, no. 2, pp. 396–413, 2010.

Bisi, A., Dada, M., Tokdar, S., "A censored-data multi-period inventory problem with newsvendor demand distributions", Manufacturing & Service Operations Management, vol. 13, no. 4, pp. 525-533, 2011.

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

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

Besbes, O., Muharremoglu, A., "On implications of demand censoring in the newsvendor problem", Management Science, vol. 59, no. 6, pp. 1407-1424, 2013.

Mansourifard, P., Krishnamachari, B., Javidi, T., "Tracking of real-valued Markovian random processes with asymmetric cost and observation", in American Control Conference (ACC). IEEE, 2015.

Tai, A. H, Ching, WK, "Recent advances on Markovian models for inventory research", International Journal of Inventory Research, vol. 3, no. 3, pp. 198-216, 2016.

Hu, J., Zhang, Ch., Zhu, Ch., "(s, S) Inventory Systems with Correlated Demands", INFORMS Journal on Computing, vol. 28, no. 4, pp. 603-611, 2016.

Alwan, L. C., Xu, M., Yao, D., Yue, X., "The dynamic newsvendor model with correlated demand", Decision Sciences, vol. 47, no. 1, pp. 11-30, 2016.

Mansourifard, P., Javidi, T., Krishnamachari, B., "Percentile Policies for Tracking of Markovian Random Processes with Asymmetric Cost and Observation", arXiv preprint arXiv:1703.01261, 2017.

Biographies

Farzaneh Mansourifard is a master student at department of engineering and technology at Alzahra university, Tehran, Iran. She earned B.S. in industrial engineering from Tabriz university, Tabriz, Iran, 2016. Her research interest is optimization with the applications in healthcare and supply chain.

Parisa Mansourifard received the B.S. and M.S. in electrical engineering from Sharif university of technology, Tehran, Iran, in 2008 and 2010 respectively. She also the M.S. in computer science and Ph.D. in electrical engineering from University of Southern California, Los Angeles, CA, USA, in 2015 and 2017, respectively. During her Ph.D. She held Viterbi Dean fellowships in 2011-2014 and AAUW dissertation completion fellowship in 2015-2016. She is currently a data scientist at Supplyframe Inc. and a part-time lecturer at University of Southern California. Her research interest is decision-making under uncertainty, machine learning and stochastic optimization.

Bhaskar Krishnamachari (M02 – SM14) received the B.E. degree in electrical engineering at The Cooper Union, New York, NY, USA, in 1998, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, USA, in 1999

and 2002, respectively. He is currently Professor and Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA. He is also the Director of the USC Viterbi School of Engineering Center on Cyber-Physical Systems and the Internet of Things. His primary research interest is in the design and analysis of algorithms, protocols, and applications for next-generation wireless networks.

Appendix A

To prove Proposition 1, we need the concavity of the value functions given in the following lemma.

Lemma 1. The expected cost-to-go accrued under action r and inventory level L, $V_t(b, L; q)$, and the value function, $V_t(b, L)$, are concave with respect to the belief vector b, i.e.

$$V_t(b, L; q) \geq \lambda V_t(b_1, L; q) + (1 - \lambda) V_t(b_2, L; q), \quad \forall r \in \mathcal{M}$$

$$V_t(b, L) \geq \lambda V_t(b_1, L) + (1 - \lambda) V_t(b_2, L), \quad \forall 0 \leq \lambda \leq 1. \qquad (17)$$

Proof of Lemma 1. We use induction to prove the concavity of $V_t(b, L; q)$ with respect to the belief vector, b, for the finite horizon. Let's assume b is a linear combination of two belief vectors b_1 and b_2, such that:

$$b = \lambda b_1 + (1 - \lambda) b_2, 0 \leq \lambda \leq 1. \qquad (18)$$

At horizon T, the immediate cost, as given in (5), is affine linear with respect to the belief vector. In other words,

$$\bar{C}(b, L; q) \geq \lambda \bar{C}(b_1, L; q) + (1 - \lambda) \bar{C}(b_2, L; q) \qquad (19)$$

which confirms the concavity of the expected cost-to-go at horizon T. Now assuming $V_{t+1}(.)$ is concave, we will consider $V_t(.)$. Using (7c) and (8) we have:

$$V_t(b, L; q) - \lambda V_t(b_1, L; q) - (1 - \lambda) V_t(b_2, L; q) = [C(b, L; q) - \lambda \bar{C}(b_1, L; q) - (1 - \lambda) \bar{C}(b_2, L; q)]$$

$$+ \sum_{i=0}^{L+q-1} [b(i) - \lambda b_1(i) - (1 - \lambda) b_2(i)] V_{t+1}(P_{i,.}, L + q - i)$$

$$+ V_{t+1}(T_{L+q}[b]P, 0) \sum_{i=L+q}^{M} b(i) - \lambda V_{t+1}(T_{L+q}[b_1]P, 0) \sum_{i=L+q}^{M} b_1(i)$$

$$- (1 - \lambda) V_{t+1}\big(T_{L+q}[b_2]P, 0\big) \sum_{i=L+q}^{M} b_2(i) \qquad (20a)$$

$$= \sum_{i=L+q}^{M} b(i) [V_{t+1}\big(T_q[b]P, 0\big) - \lambda' V_{t+1}(T_{L+q}[b_1]P, 0)$$

$$- (1 - \lambda') V_{t+1}(T_{L+q}[b_2]P, 0)] \qquad (20b)$$

where the last equality follows from (19) and $\lambda' = \lambda \frac{\sum_{i=L+q}^{M} b_1(i)}{\sum_{i=L+q}^{M} b(i)}$. Let $j \geq L + q$:

$$\lambda' T_{L+q}[b_1](j) + (1 - \lambda') T_{L+q}[b_2](j) = \frac{\lambda \sum_{i=L+q}^{M} b_1(i) T_{L+q}[b_1](j) + (1 - \lambda) \sum_{i=L+q}^{M} b_2(i) T_{L+q}[b_2](j)}{\sum_{i=L+q}^{M} b(i)}$$

$$= \frac{1}{\sum_{i=L+q}^{M} b(i)} [\lambda \sum_{i=L+q}^{M} b_1(i) \frac{b_1(j)}{\sum_{i=L+q}^{M} b_1(i)} + (1 - \lambda) \sum_{i=L+q}^{M} b_2(i) \frac{b_2(j)}{\sum_{i=L+q}^{M} b_2(i)}]$$

$$= \frac{\lambda b_1(j) + (1-\lambda) b_2(j)}{\sum_{i=L+q}^{M} b(i)} = \frac{b(j)}{\sum_{i=L+q}^{M} b(i)} = T_{L+q}[b](j). \qquad (21)$$

And for $j < L + r$, $T_{L+q}[b_1](j) + (1-\lambda')T_{L+q}[b_2](j) = 0$. Multiplying by P, we have $\lambda' T_{L+q}[b_1]P + (1-\lambda')T_{L+q}[b_2]P = T_{L+q}[b]P$. The induction step follows the concavity of $V_{t+1}(.)$.

To prove the concavity of value function, $V_t(b)$, with respect to b we use the definition of (7a) to get:

$$V_t(b, L) = \min_r V_t(b, L; q^*) = V_t(b, L; q^*) \qquad (22a)$$
$$\geq \lambda V_t(b_1, L; q^*) + (1-\lambda) V_t(b_2, L; q^*) \qquad (22b)$$
$$\geq \lambda \min_{q_1} V_t(b_1, L; q_1) + (1-\lambda) \min_{q_2} V_t(b_2, L; q_2) \qquad (22c)$$
$$= \lambda V_t(b_1, L) + (1-\lambda) V_t(b_2, L) \qquad (22d)$$

where $q^* = \arg min_q \, V_t(b;\, q)$ and (22b) is the result of the lemma for $V_t(b; q^*)$ and applying the definition, given in (7a), one more time in (22d) completes the proof.

Proof of Proposition 1. To prove this proposition, it is enough to show that,

$$V_t(b_t, L_t) - V_t^{FO}(b_t, L_t) \geq V_t\left(b_t, L_t; q_t^{opt}\right) - V_t^{FO}\left(b_t, L_t; q_t^{opt}\right) \geq 0 \qquad (23)$$

for $q_t^{opt} = \arg min_r \, V_t(b_t; q)$. First, the cost-to-go function of FO case can be computed as:

$$V_t^{FO}(b_t; q) = \bar{C}(b_t; q) + \sum_{i=0}^{M} b_t(i) V_{t+1}^{FO}\left(P_{y,.}\right), \qquad (24)$$

Now to proof the proposition we use induction. Fist, at $t = T$ we have:

$$V_T\left(b_T, L_T; q_T^{opt}\right) - V_T^{FO}\left(b_T, L_T; q_T^{opt}\right) = \bar{C}\left(b_T, L_T; q_T^{opt}\right) - \bar{C}\left(b_T, L_T; q_T^{opt}\right) = 0 \qquad (25)$$

Now assuming (14) is true at time steps t + 1 onwards, we should prove it for time t.

$$V_t(b_t, L_t) - V_t^{FO}(b_t, L_t) \geq [\bar{C}\left(b_t, L_t; q_t^{opt}\right) + \sum_{i=0}^{L_t + q_t^{opt} - 1} b_t(i) V_{t+1}(P_{i,.}, L_t + q_t^{opt} - i)$$
$$+ \sum_{i=L_t+q_t^{opt}}^{M} b_t(i) V_{t+1}(T_{L_t+q_t^{opt}}[b_t]P, 0)]$$
$$-[\bar{C}\left(b_t, L_t; q_t^{opt}\right) + \sum_{i=0}^{L_t + q_t^{opt} - 1} b_t(i) v_{t+1}^{FO}(P_{i,.}, L_t + q^{opt}(b_t) - i)$$
$$+ \sum_{i=L_t+q_t^{opt}}^{M} b_t(i) V_{t+1}^{FO}(P_{i,.}, 0)], \qquad (26)$$

Thus,

$$V_t(b_t, L_t) - V_t^{FO}(b_t, L_t) \geq \sum_{i=0}^{L_t + q_t^{opt} - 1} b_t(i) V_{t+1}\left(P_{i,.}, L_t + q_t^{opt} - i\right) - V_{t+1}^{FO}(P_{i,.}, L_t + q_t^{opt} - i)]$$
$$+ \sum_{i=L_t+q_t^{opt}}^{M} b_t(i)[V_{t+1}\left(T_{L_t+q_t^{opt}}[b_t]P, 0\right) - V_{t+1}^{FO}\left(P_{i,.}, 0\right)] \qquad (27)$$

The first term is greater than or equal to zero based on the concavity at t + 1. We use the concavity of the value function to get the following inequality:

$$V_{t+1}\left(T_{L_t+q_t^{opt}}[b_t]P,0\right) \geq \frac{\sum_{i=L_t+q_t^{opt}}^{M} b_t(i)V_{t+1}\left(P_{i,.},0\right)}{\sum_{j=L_t+q_t^{opt}}^{M} b_t(j)}. \quad (28)$$

Therefore, by applying (28) and the induction assumption at t + 1, we have:

$$\sum_{i=L_t+q_t^{opt}}^{M} b_t(i)[V_{t+1}\left(T_{L_t+q_t^{opt}}[b_t]P,0\right) - V_{t+1}^{FO}\left(P_{i,.},0\right)] \geq \sum_{i=L_t+q_t^{opt}}^{M} b_t(i)\left[V_{t+1}\left(P_{i,.},0\right) - V_{t+1}^{FO}\left(P_{i,.},0\right)\right] \geq 0 \quad (29)$$

thus, (27) is greater than or equal to zero. This completes the proof.