Structural Results for Average-Cost Inventory Models with Markov-Modulated Demand and Partial Information

Harun Avcı, Kağan Gökbayrak, Emre Nadar (corresponding author)
Department of Industrial Engineering, Bilkent University, 06800 Ankara, Turkey
{harun.avci@bilkent.edu.tr, kgokbayr@bilkent.edu.tr, emre.nadar@bilkent.edu.tr}

We consider a discrete-time infinite-horizon inventory system with non-stationary demand, full backlogging, and deterministic replenishment lead time. Demand arrives according to a probability distribution conditional on the state of the world that undergoes Markovian transitions over time. But the actual state of the world can only be imperfectly estimated based on past demand data. We model the inventory replenishment problem for this system as a Markov decision process (MDP) with an *uncountable* state space consisting of both the inventory position and the most recent *belief*, a conditional probability mass function, about the actual state of the world. Assuming that the state of the world evolves as an ergodic Markov chain, using the vanishing discount method along with a coupling argument, we prove the existence of an optimal average cost that is independent of the initial system state. For our linear cost structure, we also establish the average-cost optimality of a belief-dependent base-stock policy. We then discretize the uncountable belief space into a regular grid and observe that the average cost under our discretization converges to the optimal average cost as the number of grid points grows large. Finally, we conduct numerical experiments to evaluate the use of a myopic belief-dependent base-stock policy as a heuristic for our MDP with the uncountable state space. On a test bed of 108 instances, the average cost obtained from the myopic policy deviates by no more than a few percent from the best lower bound on the optimal average cost obtained from our discretization.

Key words: inventory control; Markov-modulated demand; partial information; long-run average cost

History: Received: September 2018; Accepted: July 2019 by Jayashankar Swaminathan, after 2 revisions.

1. Introduction

Companies often face non-stationary demand that is driven by dynamic environmental factors, such as fluctuating economic and/or market conditions (Shang 2012, Kesavan and Kushwaha 2014, and Hu et al. 2016). Associating a specific demand distribution with each state, Markov chains provide an elegant mathematical framework for incorporating non-stationary demand into inventory models. In this framework, the probability distribution of demand evolves over time according to a Markov chain whose state variable captures all the relevant information about environmental factors to represent the *demand state*. But the current demand state is often unobservable. The true demand state is unknown to a manufacturer until she notices the introduction, change in relative price, or end-of-life decision of a competing product, or the change in purchasing power or interest of customers. In a supply chain setting, it may also be unknown to upstream stages if the information provided by downstream stages is limited or its credibility cannot be guaranteed

(Özer et al. 2011, Shamir and Shin 2016, and Spiliotopoulou et al. 2016). Limited information sharing may even arise within a manufacturing firm: The authors encountered a situation in a furniture company where the sales department did not communicate the promotion periods to the production department. This company thus observed raw material stock-outs due to high demand during the promotion periods, but also suffered from excess inventory in the subsequent periods since it raised its order quantities by misinterpreting this short-lived upward shift in demand.

Only a few researchers have considered Markov-modulated demand under *partial* information about the demand state. These researchers have focused only on finite-horizon total-cost and infinite-horizon discounted-cost inventory systems with bounded or Poisson demand (Treharne and Sox 2002, Bayraktar and Ludkovski 2010, and Malladi et al. 2019). To our knowledge, in the literature dealing with Markov-modulated demand and partial information, no one has studied the infinite-horizon average-cost inventory systems. One reason for this gap in the literature is the notorious difficulty of the resulting partially observed Markov decision processes (POMDPs) under the average cost criterion (Ding et al. 2002 and Chapter 5 in Bertsekas 2012). Our study is the first attempt to fill this gap: We establish structural results for an *average*-cost inventory system with Markov-modulated demand and partial information.

Specifically, we study the inventory replenishment problem for a single-item discrete-time system with full backlogging and non-stationary demand that arrives according to one of a finite number of probability distributions in each time period. The probability distributions undergo Markovian transitions between time periods. The state of the underlying Markov chain, i.e., the demand state, is only partially observable based on past demand data. The infinite-horizon discounted-cost problem for this system can be modeled as a POMDP with an information vector that contains all past demand observations and the belief about the initial demand state. The demand state belief in any period can be specified as a probability distribution over the set of demand states that forms a sufficient statistic for the entire demand history and possesses the Markovian property. The belief evolves over time, as new demand observations become available, according to the Bayes' formula. To leverage this Markovian structure of the belief, we formulate the infinite-horizon discounted-cost problem as an MDP with a state space consisting of the inventory position and the belief about the current demand state. (See Sandıkçı 2010 for details on reduction of a POMDP to an MDP.)

Research on inventory management with demand information updating can be classified into three groups: The first group employs time series to model the demand (e.g., Johnson and Thompson 1975 and Miller 1986). The second group uses the martingale model of forecast evolution (e.g., Graves et al. 1986, Heath and Jackson 1994, Kaminsky and Swaminathan 2001, 2004). The last group exploits Bayesian updating mechanisms for stationary demand with unknown parameters under backlogging (e.g., Scarf 1959, 1960, Karlin 1960, Iglehart 1964, and Azoury 1985), for censored demand under unobservable lost sales (e.g., Lariviere and Porteus 1999, Ding et al. 2002, Heese and Swaminathan 2010, and Chen and Mersereau 2015), and for non-stationary demand with partially observed demand states (e.g., Treharne and Sox 2002, Arifoğlu and Özekici 2010, 2011, and Malladi et al. 2019). Our study falls into the last group.

For our linear cost structure, we establish the optimality of a belief-dependent base-stock policy in the discounted-cost case, deriving theoretical bounds on the optimal base-stock levels (Proposition 1). However, this result is not immediate in the average-cost case: The state space of our MDP with Bayesian updating is uncountable due to the presence of the demand state belief. This hinders straightforward application of the standard vanishing discount method in the average-cost analysis of our problem. It is even unclear whether there exists an optimal average cost that is independent of the initial state of our MDP. Assuming that the Markov chain governing the demand state transitions is ergodic (Assumption 1), we are able to adapt the vanishing discount method to our MDP via a coupling argument that was not studied in the extant inventory literature to our knowledge. This enables us to prove that (i) there exists an optimal average cost independent of the initial state of our MDP, (ii) the average-cost optimality equation holds, and (iii) the belief-dependent base-stock policy is optimal in the average-cost case (Theorem 1).

Since the state space of our MDP is uncountable, finding an exact solution for the average-cost optimality equation (and calculating the base-stock levels) is a computational challenge (Zhou and Hansen 2001 and Saldı et al. 2017). As an approximation, we discretize our belief space via the regular grid approach in Lovejoy (1991). The average cost under this approximation is a lower bound on the optimal average cost (Yu and Bertsekas 2004). This bound converges to the optimal average cost as the number of grid points goes to infinity. Compared to this lower bound, we numerically evaluate the cost performance of a *myopic* belief-dependent base-stock policy as a heuristic replenishment policy for our average-cost problem with uncountable state space.

Myopic base-stock policies can be easily implemented in practice. These policies were also shown to be optimal for several inventory models in the case of stationary demand (Veinott 1965 and Lovejoy 1990) and in the case of non-stationary demand under certain conditions (Johnson and

Thompson 1975 and Lovejoy 1992). For the finite-horizon total-cost version of our problem, however, previous research has found that the total cost under the myopic belief-dependent base-stock policy may significantly deviate from the optimal total cost (see Treharne and Sox 2002): On a test bed of 252 instances, for the myopic policy, the average optimality gap is 5.19% and the largest optimality gap is 44.84%. Our numerical experiments reveal that the myopic policy performs significantly better in our average-cost problem than in the finite-horizon total-cost problem: On the same test bed, the average cost under the myopic policy is only 0.41% on average, and no more than 3.61%, higher than the best lower bound on the optimal average cost that can be obtained from our approximation. It is also noteworthy that the myopic policy computations are instantaneous.

The rest of this paper is organized as follows: Section 2 reviews the related literature. Section 3 formulates our problem. Section 4 presents our structural results for both discounted-cost and average-cost problems. Section 5 offers a discretization scheme for the calculation of a lower bound on the optimal average cost. Section 6 presents our numerical results and Section 7 concludes.

2. Related Literature

Most classical inventory models assume that demand is independent of environmental factors other than time (Chapter 1 in Beyer et al. 2010). There is also a growing body of literature that models non-stationary demand (due to environmental factors) as a Markov-modulated process: Song and Zipkin (1993) consider an inventory system with Markov-modulated Poisson demand, full backlogging, ordered stochastic replenishment lead times, and fixed and linear variable ordering costs. They establish the optimality of a state-dependent (s, S) policy in the discounted cost problem. In the case of zero lead time, Sethi and Cheng (1997) generalize the optimality of state-dependent (s, S) policies to inventory systems with more general Markov-modulated demand, full backlogging, and fixed and linear variable ordering costs. Applying the vanishing discount method to the infinite-horizon discounted-cost problem in Sethi and Cheng (1997), Beyer and Sethi (1997) extend the optimality of state-dependent (s, S) policies to the infinite-horizon average-cost problem. Using the vanishing discount method, Huh et al. (2011) partially characterize the optimal policy structures for several different single-stage inventory models with Markov-modulated demand and capacity.

All of the papers above assume that the current state of the Markov-modulated process is perfectly observed by the controller and thus the true demand distribution is always known. Several other papers have significantly relaxed this assumption: Treharne and Sox (2002) consider discrete-time inventory systems in which the demand state can only be partially observed through the past demand data. They study the finite-horizon total-cost problem with bounded demand, full backlogging, deterministic lead time, and linear variable ordering costs. They establish the optimality of a base-stock policy with the base-stock levels that depend on the most recent belief about the actual demand state. Arifoğlu and Özekici (2010) consider discrete-time inventory systems with random yield, finite capacity, full backlogging, zero lead time, and fixed and linear variable ordering costs. The demand state is partially revealed via an observation process that is different from the past demand data. The observation process takes values from a finite set whereas the demand is real valued. They prove the optimality of belief-dependent (s, S) policies in the discounted-cost problem. Bayraktar and Ludkovski (2010) consider continuous-time inventory systems with Markov-modulated Poisson demand, zero lead time, and fixed and linear variable ordering costs. The demand state is partially observed through the past demand data. They characterize the optimal policy structure in both cases of backlogging and lost sales. Malladi et al. (2019) consider discrete-time inventory systems with finite demand, full backlogging, zero lead time, fixed and linear variable ordering costs. The demand state is partially revealed via an observation process that contains the past demand data. All these papers incorporate partial observations into their inventory models via Bayesian updating mechanisms in finite-horizon total-cost or infinite-horizon discounted-cost problems. In this study, however, we focus on the infinite-horizon average-cost problem. In addition, we allow for more general demand distributions by only requiring the second moment of demand to be finite (in each demand state).

For the infinite-horizon average-cost problems with uncountable state spaces, the optimal average cost may depend on the initial state. And when it is independent of the initial state, an optimal stationary policy need not exist (Chapter 5 in Bertsekas 2012). The vanishing discount method can be used to prove the existence of a constant optimal average cost that is independent of the initial state. Using this method, Ross (1968) shows that the uniform boundedness and equicontinuity of the differential discounted cost function ensures the existence of an optimal average cost. Beyer and Sethi (1997) establish the uniform boundedness and equicontinuity of the differential discounted cost function for inventory models in which the *perfectly* observed demand state evolves over time according to an *irreducible* Markov chain. Using a coupling argument to obtain certain bounds on the discounted cost function, Borkar (2000) proves the uniform boundedness and equicontinuity of the differential discounted cost function for controlled Markov chains with partial observations when the underlying Markov chain is ergodic. In this study, we apply the vanishing discount method together with the coupling argument of Borkar (2000) to our inventory system, which enables us to show the existence of a constant optimal average cost for this system.

Since solving the average-cost optimality equation on an uncountable state space is infeasible, previous work has developed discretization schemes for approximate solutions. Lovejoy (1991) discretizes the uncountable state space into a regular grid with the concept of "triangulation." Zhou and Hansen (2001) improve Lovejoy's result by introducing a variable-resolution regular grid. Both papers establish a lower bound for discounted-cost problems modeled as POMDPs. Yu and Bertsekas (2004) present a lower approximation approach for both discounted-cost and average-cost problems modeled as POMDPs. There are also papers that approximate the average cost for MDPs with uncountable state space; see, for instance, Feinberg et al. (2012) and Saldı et al. (2017). In this study, we adopt the discretization schemes in Lovejoy (1991) and Yu and Bertsekas (2004), which enable us to obtain a sufficiently tight lower bound on the optimal average cost.

We thus contribute to the inventory literature in several important ways:

- To our knowledge, we are the first to study the inventory replenishment problem with Markov-modulated demand and partial demand state information in the average cost case.
- In the literature, Treharne and Sox (2002) have shown the optimality of a belief-dependent base-stock policy in the finite-horizon problem with bounded demand. We generalize this structural result to the infinite-horizon discounted-cost problem with possibly unbounded demand. We also derive theoretical bounds on the optimal base-stock levels.
- We are the first to apply the coupling argument of Borkar (2000) along with the vanishing discount method to the inventory replenishment problem. With this approach, we prove the existence of an optimal average cost independent of the initial state of our MDP defined on an uncountable state space and extend the optimality of the belief-dependent base-stock policy to the average-cost problem. Our approach may also conceivably be useful in other operations management problems modeled as POMDPs under the average-cost criterion.
- We propose the use of a myopic belief-dependent base-stock policy as a heuristic. We numerically examine its performance by calculating a sufficiently tight lower bound on the optimal average cost, obtained from the discretized version of our problem as in Lovejoy (1991) and Yu and Bertsekas (2004), and by comparing this bound to the cost under the myopic policy.

3. Problem Formulation

We study the inventory replenishment problem for a single-item discrete-time system over an infinite horizon. Demand in each period arrives according to a probability distribution conditional on the state of world (e.g., economy or market) that undergoes Markovian transitions over time. The demand state in period t, d_t, takes a value from a finite set $\mathcal{N} := \{1, 2, ..., N\}$, $\forall t \in \mathbb{Z}_+ := \{1, 2, ...\}$. We model the demand state process $\{d_t\}_{t \in \mathbb{Z}_+}$ as a finite-state Markov chain with an $N \times N$ stationary transition matrix $P = \{p_{ij}\}$ where $p_{ij} := \mathbb{P}\{d_{t+1} = j | d_t = i\}$, $\forall t \in \mathbb{Z}_+$. Demand in period t, w_t, is a non-negative discrete random variable. We denote by $r_i(\cdot)$ the probability mass function of w_t for a given $d_t = i$, i.e., $r_i(k) := \mathbb{P}\{w_t = k | d_t = i\}$. The values of p_{ij} and $r_i(\cdot)$ can be estimated based on past demand data by employing the Baum-Welch algorithm from the literature that builds upon the maximum likelihood estimation procedure. We refer the reader to Rabiner (1989) for a detailed description of this algorithm. We assume that there exist a finite μ and a finite ζ such that $\mu_i = \mathbb{E}[w_t | d_t = i] \leq \mu$ and $\zeta_i = \mathbb{E}[w_t^2 | d_t = i] \leq \zeta$, $\forall i \in \mathcal{N}$. We also assume that there exists an $i \in \mathcal{N}$ such that $\mu_i > 0$. The latter assumption is violated if and only if the demand is always zero throughout the infinite horizon.

The demand state d_t, $t \in \mathbb{Z}_+$, can only be partially observed based on the initial state belief and the realized demand values prior to period t. We define the state belief in any period, which is also known as the "conditional state distribution" in the literature (see Fernández-Gaucherand et al. 1991), as an N-dimensional vector consisting of the apriori probabilities of being in each demand state conditioned on the history composed of the initial state belief and all past demand observations. Therefore, the belief in period $t > 1$, $\pi^t = [\pi_1^t, \ldots, \pi_N^t]$, can be formulated as $\pi_i^t(\pi, \omega) := \mathbb{P}\{d_t = i | \pi^1 = \pi, \omega^{t-1} = \omega\}$, $\forall i \in \mathcal{N}$, where $\pi^1 = [\pi_1^1, \ldots, \pi_N^1]$ is the initial state belief and $\omega^{t-1} = (w_1, ..., w_{t-1})$ is the demand history. Note that $\pi_i^1 = \mathbb{P}\{d_1 = i\}$, $\forall i \in \mathcal{N}$. We denote by $\hat{r}_\pi(\cdot)$ the probability mass function of w_t for a given $\pi^t = \pi$. Thus:

$$\hat{r}_\pi(k) = \mathbb{P}\{w_t = k | \pi^t = \pi\} = \sum_{i \in \mathcal{N}} \mathbb{P}\{w_t = k | \pi^t = \pi, d_t = i\} \mathbb{P}\{d_t = i | \pi^t = \pi\} = \sum_{i \in \mathcal{N}} r_i(k) \pi_i.$$

Also, note that $\mu_\pi = \mathbb{E}[w_t | \pi^t = \pi] = \sum_{i \in \mathcal{N}} \pi_i \mu_i$ and $\zeta_\pi = \mathbb{E}[w_t^2 | \pi^t = \pi] = \sum_{i \in \mathcal{N}} \pi_i \zeta_i$.

All unmet demand is backlogged. The replenishment order placed at the beginning of period t is received at the beginning of period $t + l$, where $l \in \{0, 1, \ldots\}$ is constant, $\forall t \in \mathbb{Z}_+$. As we allow for non-zero replenishment lead times, we define the inventory position as the number of items on hand plus the number of items on order minus the number of backlogged demands, and include it in the state description of our MDP. We denote the inventory position at the beginning of period

t by $y_t \in \mathbb{Z}$, and the replenishment order quantity at the beginning of period t by $u_t \in \mathbb{Z}_+ \cup \{0\}$, $\forall t \in \mathbb{Z}_+$. For an initial inventory position y_1, the inventory position evolves over time as follows:

$$y_{t+1} = y_t + u_t - w_t, \quad \forall t \in \mathbb{Z}_+. \tag{1}$$

Since the state belief π^t forms a sufficient statistic for the information collected up to period t (see Rhenius 1974), we also include π^t in the state description of our MDP. Given the current belief $\pi^t = \pi$ and the current demand realization $w_t = w$, the next belief π^{t+1} can be calculated as follows:

$$\pi_i^{t+1} = \frac{\sum_{j \in \mathcal{N}} p_{ji}\ r_j(w)\ \pi_j}{\sum_{j' \in \mathcal{N}} r_{j'}(w)\ \pi_{j'}} := T_i(\pi, w), \quad \forall t \in \mathbb{Z}_+, \forall i \in \mathcal{N}. \tag{2}$$

Let $\Pi := \left\{\pi \in [0,1]^N \ : \ \sum_{i \in \mathcal{N}} \pi_i = 1\right\}$ be the continuous space of all possible beliefs. We define $T : \Pi \times (\mathbb{Z}_+ \cup \{0\}) \to \Pi$ as the one-period belief update function given by $T(\cdot,\cdot) = [T_1(\cdot,\cdot), \ldots, T_N(\cdot,\cdot)] \in \Pi$ (see Treharne and Sox 2002 and Chapter 4 in Bertsekas 2017 for similar belief updates).

There are two types of costs in our inventory model: The ordering cost in period t is linear in the order quantity and is given by cu_t, where c is the unit ordering cost. The single-period expected inventory cost in period $t+l$ is piecewise linear and is given by

$$g(\pi^t, y_t + u_t) = \mathbb{E}\left[\max\left\{h\left(y_t + u_t - \sum_{n=0}^{l} w_{t+n}\right), b\left(-y_t - u_t + \sum_{n=0}^{l} w_{t+n}\right)\right\}\middle|\pi^t\right],$$

where b and h are the unit shortage and holding costs per period, respectively. Note that the conditional $(l+1)$-period demand distribution can be calculated as follows:

$$\mathbb{P}\left\{\sum_{n=0}^{l} w_{t+n} = k \middle| \pi^t\right\} = \sum_{k_1=0}^{k} \sum_{k_2=0}^{k-k_1} \cdots \sum_{k_l=0}^{k-\sum_{j=1}^{l-1} k_j} \hat{r}_{\pi^t}(k_1)\hat{r}_{\pi^{t+1}}(k_2)\cdots\hat{r}_{\pi^{t+l-1}}(k_l)\hat{r}_{\pi^{t+l}}\left(k - \sum_{j=1}^{l} k_j\right)$$

where $\pi^{t+1} = T(\pi^t, k_1), \ldots, \pi^{t+l} = T(\pi^{t+l-1}, k_l)$.

For any initial belief $\pi^1 = \pi \in \Pi$ and any initial inventory position $y_1 = y \in \mathbb{Z}$, the expected long-run average cost per period under a replenishment policy with order quantities $U = (u_1, u_2, ...)$, $u_t \geq 0$, $t \in \mathbb{Z}_+$, can be written as

$$J^U(\pi, y) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} [cu_t + g(\pi^t, y_t + u_t)] \middle| \pi^1 = \pi, y_1 = y\right] \text{ subject to (1) and (2).}$$

The objective is to determine the replenishment policy that minimizes the expected long-run average cost per period. Since the replenishment decisions cannot influence the inventory levels up to period l, we exclude the holding and shortage costs incurred up to period l from our cost formulation. However, the belief updating process begins with the initial demand observation at the end of the first period. In this formulation we allow the order quantity to depend on the state of the system in each period. For notational convenience, however, we suppress the dependency of u_t on (π^t, y_t). In Section 4, we prove that there exists a replenishment policy with order quantities $U^* = (u_1^*, u_2^*, \ldots) \in \mathcal{U}$, where $\mathcal{U} = \{(u_1, u_2, \ldots) | u_t \in \mathbb{Z}_+ \cup \{0\}\}$, and a constant λ^*, which is independent of π and y, such that $\lambda^* = J^{U^*}(\pi, y) \leq J^U(\pi, y)$, $\forall U \in \mathcal{U}$, $\forall \pi \in \Pi$, and $\forall y \in \mathbb{Z}$.

4. Analytical Results

In this section, first, we provide structural results for the discounted-cost problem that we will utilize in our average-cost analysis. Then, we employ the vanishing discount method along with a coupling argument for arbitrarily different initial demand states to prove that the average-cost optimality equation holds, and use this equation to characterize the optimal policy structure for our average-cost problem. We refer the reader to Chapter 5 in Beyer et al. (2010), Chapter 5 in Bertsekas (2012), and Chapter 8 in Puterman (2014) for details on the vanishing discount method.

4.1. The Discounted-Cost Problem

Let $\alpha \in (0,1)$ denote the discount factor. For any initial state $(\pi, y) \in \Pi \times \mathbb{Z}$, the optimal expected total discounted cost over an infinite horizon can be defined as

$$v_\alpha(\pi, y) = \inf_{U \in \mathcal{U}} J_\alpha^U(\pi, y)$$

where $J_\alpha^U(\pi, y)$ is the expected total discounted cost for the initial state (π, y) under a replenishment policy with order quantities $U = (u_1, u_2, \ldots)$, i.e.,

$$J_\alpha^U(\pi, y) = \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=1}^{T} \alpha^{t-1}[cu_t + \alpha^l g(\pi^t, y_t + u_t)] \middle| \pi^1 = \pi, y_1 = y\right].$$

Following Proposition 4.1.9 in Bertsekas (2012), we verify that the optimal cost function v_α satisfies

$$v_\alpha(\pi, y) = \min_{u \geq 0}\left\{cu + \alpha^l g(\pi, y+u) + \alpha \sum_{w=0}^{\infty} v_\alpha(T(\pi, w), y+u-w)\hat{r}_\pi(w)\right\}, \ \forall \pi \in \Pi, \ \forall y \in \mathbb{Z}. \tag{3}$$

(See Treharne and Sox 2002 for a similar formulation on the finite-horizon total-cost problem with finite demand.) We assume that $\alpha^l b > c$. Note that if $\alpha^l b$ were less than c, it would never be optimal to place an order in an $(l+1)$-period problem. This assumption is standard in the inventory literature; see, for instance, Sethi and Cheng (1997), Arifoğlu and Özekici (2010), and Chapter 3 in Bertsekas (2017). For the discounted-cost problem in (3), Proposition 1 shows that a belief-dependent base-stock policy is optimal and the optimal belief-dependent base-stock levels S_α^π are bounded between 0 and $\frac{\alpha^l \mu(b+h)(l+1)}{\alpha^l h + (1-\alpha)c}$, $\forall \pi \in \Pi$, $\forall \alpha \in (0,1)$.

Proposition 1. *For any $\alpha \in (0,1)$, the optimal stationary inventory replenishment policy is a belief-dependent base-stock policy with base-stock levels S_α^π such that the optimal order quantity in state (π, y) is $S_\alpha^\pi - y$ if $y < S_\alpha^\pi$, and zero otherwise. Furthermore, the optimal belief-dependent base-stock levels S_α^π, $\forall \pi \in \Pi$, satisfy (i) $S_\alpha^\pi \leq M$ where $M := \frac{\alpha^l \mu(b+h)(l+1)}{\alpha^l h + (1-\alpha)c}$ and (ii) $S_\alpha^\pi \geq 0$.*

Proof. We will prove that $v_\alpha(\pi, y)$ is discrete-convex in y, i.e., $v_\alpha(\pi, y+1) - v_\alpha(\pi, y) \geq v_\alpha(\pi, y) - v_\alpha(\pi, y-1)$, $\forall \pi \in \Pi$, $\forall y \in \mathbb{Z}$. With this property, we are able to characterize the optimal policy structure. We consider the value iteration algorithm that can be used to calculate $v_\alpha(\cdot, \cdot)$: Let $v_\alpha^t(\cdot, \cdot)$ denote the cost function at iteration t of the value iteration algorithm. Letting $z = y + u$, we obtain $v_\alpha^{t+1}(\pi, y) = -cy + \min_{z \geq y}\{G_\alpha^t(\pi, z)\}$ where

$$G_\alpha^t(\pi, z) = cz + \alpha^l g(\pi, z) + \alpha \sum_{w=0}^{\infty} v_\alpha^t(T(\pi, w), z - w)\hat{r}_\pi(w).$$

Since $g(\pi, z) \geq 0$ and $G_\alpha^t(\pi, z) \geq cz + \alpha^l g(\pi, z)$, $\lim_{z\to+\infty} G_\alpha^t(\pi, z) \geq \lim_{z\to+\infty}\{cz + \alpha^l g(\pi, z)\} = \infty$ and $\lim_{z\to-\infty} G_\alpha^t(\pi, z) \geq \lim_{z\to-\infty}\{cz + \alpha^l g(\pi, z)\} \geq \lim_{z\to-\infty}\{(c - \alpha^l b)z\} = \infty$ (recall $\alpha^l b > c$). These results also hold when $c = 0$ by definition of $g(\pi, z)$ and our assumption of $\mathbb{E}[w_t] \leq \mu$, $\forall t \in \mathbb{Z}_+$. We assume that $v_\alpha^0(\cdot, \cdot)$ is the zero function. Thus, following Proposition 4.1.9 in Bertsekas (2012), we verify $\lim_{t\to\infty} v_\alpha^t(\pi, y) = v_\alpha(\pi, y)$.

Note that $v_\alpha^0(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$. Also, note that $G_\alpha^0(\pi, 0) = \alpha^l g(\pi, 0) \leq \alpha^l b\mu(l+1) < \infty$, $\forall \pi \in \Pi$. Assuming that $v_\alpha^t(\pi, y)$ is discrete-convex in y and $G_\alpha^t(\pi, 0)$ is finite, $\forall \pi \in \Pi$, we will show that $v_\alpha^{t+1}(\pi, y)$ is discrete-convex in y and $G_\alpha^{t+1}(\pi, 0)$ is finite, $\forall \pi \in \Pi$. First, we show that $v_\alpha^{t+1}(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$. It is easy to verify that $cz + \alpha^l g(\pi, z)$ is discrete-convex in z. Thus, as we assume $v_\alpha^t(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$, and since $G_\alpha^t(\cdot, \cdot)$ is a sum of discrete-convex functions, $G_\alpha^t(\pi, z)$ is discrete convex in z, $\forall \pi \in \Pi$. Hence, as we assume that $G_\alpha^t(\pi, 0) < \infty$, and since $\lim_{z\to+\infty} G_\alpha^t(\pi, z) = \lim_{z\to-\infty} G_\alpha^t(\pi, z) = \infty$, $\forall \pi \in \Pi$, there exists a finite global minima $S_\alpha^{\pi,t}$ such that $S_\alpha^{\pi,t} = \arg\min_{z\in\mathbb{Z}}\{G_\alpha^t(\pi, z)\}$, $\forall \pi \in \Pi$. This implies that

$$\min_{z \geq y} G_\alpha^t(\pi, z) = \begin{cases} G_\alpha^t(\pi, S_\alpha^{\pi,t}) & \text{if } y < S_\alpha^{\pi,t}, \\ G_\alpha^t(\pi, y) & \text{if } y \geq S_\alpha^{\pi,t}. \end{cases}$$

In order to show that $v_\alpha^{t+1}(\pi, y)$ is discrete convex in y, we need to consider three different cases depending on the relationship between $S_\alpha^{\pi,t}$ and y:

(1) If $y < S_\alpha^{\pi,t}$, we have $v_\alpha^{t+1}(\pi, y+1) = -c(y+1) + \min_{z\geq y+1}\{G_\alpha^t(\pi, z)\} = -c(y+1) + G_\alpha^t(\pi, S_\alpha^{\pi,t}), v_\alpha^{t+1}(\pi, y) = -cy + \min_{z\geq y}\{G_\alpha^t(\pi, z)\} = -cy + G_\alpha^t(\pi, S_\alpha^{\pi,t})$, and $v_\alpha^{t+1}(\pi, y-1) = -c(y-1) + \min_{z\geq y-1}\{G_\alpha^t(\pi, z)\} = -c(y-1) + G_\alpha^t(\pi, S_\alpha^{\pi,t})$. Hence $v_\alpha^{t+1}(\pi, y+1) - v_\alpha^{t+1}(\pi, y) = -c = v_\alpha^{t+1}(\pi, y) - v_\alpha^{t+1}(\pi, y-1)$.

(2) If $y = S_\alpha^{\pi,t}$, we have $v_\alpha^{t+1}(\pi, y+1) = -c(y+1) + G_\alpha^t(\pi, S_\alpha^{\pi,t}+1), v_\alpha^{t+1}(\pi, y) = -cy + G_\alpha^t(\pi, S_\alpha^{\pi,t})$, and $v_\alpha^{t+1}(\pi, y-1) = -c(y-1) + G_\alpha^t(\pi, S_\alpha^{\pi,t})$. Since $S_\alpha^{\pi,t}$ is the global minima, $v_\alpha^{t+1}(\pi, y+1) - v_\alpha^{t+1}(\pi, y) = -c + G_\alpha^t(\pi, S_\alpha^{\pi,t}+1) - G_\alpha^t(\pi, S_\alpha^{\pi,t}) \geq -c = v_\alpha^{t+1}(\pi, y) - v_\alpha^{t+1}(\pi, y-1)$.

(3) If $y > S_\alpha^{\pi,t}$, we have $v_\alpha^{t+1}(\pi, y+1) = -c(y+1) + G_\alpha^t(\pi, y+1), v_\alpha^{t+1}(\pi, y) = -cy + G_\alpha^t(\pi, y)$, and $v_\alpha^{t+1}(\pi, y-1) = -c(y-1) + G_\alpha^t(\pi, y-1)$. By discrete-convexity of $G_\alpha^t(\pi, z)$, $v_\alpha^{t+1}(\pi, y+1) - v_\alpha^{t+1}(\pi, y) = -c + G_\alpha^t(\pi, y+1) - G_\alpha^t(\pi, y) \geq -c + G_\alpha^t(\pi, y) - G_\alpha^t(\pi, y-1) = v_\alpha^{t+1}(\pi, y) - v_\alpha^{t+1}(\pi, y-1)$.

Hence $v_\alpha^{t+1}(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$. Next, we show that $G_\alpha^{t+1}(\pi, 0)$ is finite, $\forall \pi \in \Pi$. As we assume that $G_\alpha^t(\pi, 0) < \infty$, and since $v_\alpha^{t+1}(\pi, y) \leq -cy + G_\alpha^t(\pi, 0)$, $\forall y \leq 0$, we have $\sum_{w=0}^{\infty} v_\alpha^{t+1}(T(\pi, w), -w)\hat{r}_\pi(w) \leq \sum_{w=0}^{\infty}[cw + G_\alpha^t(T(\pi, w), 0)]\hat{r}_\pi(w) = c\mu_\pi + \sum_{w=0}^{\infty} G_\alpha^t(T(\pi, w), 0)\hat{r}_\pi(w) < \infty$. Hence $G_\alpha^{t+1}(\pi, 0) < \infty$, $\forall \in \Pi$.

Since $\lim_{t\to\infty} v_\alpha^t(\pi, y) = v_\alpha(\pi, y)$, $v_\alpha(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$. Let $G_\alpha(\pi, z) = cz + \alpha^l g(\pi, z) + \alpha \sum_{w=0}^{\infty} v_\alpha(T(\pi, w), z-w)\hat{r}_\pi(w)$. Since $v_\alpha(\pi, y)$ is discrete-convex in y, $\forall \pi \in \Pi$, $G_\alpha(\pi, z)$ is discrete-convex in z, $\forall \pi \in \Pi$. Also, note that $\lim_{z\to+\infty} G_\alpha(\pi, z) = \lim_{z\to-\infty} G_\alpha(\pi, z) = \infty$ and $\lim_{t\to\infty} G_\alpha^t(\pi, 0) = G_\alpha(\pi, 0) < \infty$, $\forall \pi \in \Pi$. Therefore a belief-dependent base-stock policy with base-stock levels S_α^π is optimal. We next prove (i) and (ii):

(i) For any $\alpha \in (0,1)$, let $U = (u_1, u_2, \ldots)$ represent the order quantities under the optimal belief-dependent base-stock levels S_α^π, $\forall \pi \in \Pi$. Suppose that $\exists \pi \in \Pi$ such that $S_\alpha^\pi > M$. Now consider all sample paths that start with $y_1 = y$ for some finite $y < S_\alpha^\pi$ and $\pi^1 = \pi$ where $S_\alpha^\pi > M$. We now construct an alternative policy with order quantities $\widetilde{U} = (\widetilde{u}_1, \widetilde{u}_2, \ldots)$ such that

$$\widetilde{u}_t = \begin{cases} u_1 - 1 & \text{if } t = 1, \\ u_2 + 1 & \text{if } t = 2, \\ u_t & \text{otherwise.} \end{cases}$$

The inventory position plus the order quantity in period t under the alternative policy is

$$\widetilde{y}_t + \widetilde{u}_t = \begin{cases} y_1 + u_1 - 1 & \text{if } t = 1, \\ y_t + u_t & \text{if } t > 1. \end{cases}$$

Hence:

$$\begin{aligned} J_\alpha^{\widetilde{U}}(\pi, y) - J_\alpha^U(\pi, y) &= \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1}[c\widetilde{u}_t + \alpha^l g(\pi^t, \widetilde{y}_t + \widetilde{u}_t) - cu_t - \alpha^l g(\pi^t, y_t + u_t)] \middle| \pi^1 = \pi, y_1 = \widetilde{y}_1 = y\right] \\ &= \alpha^l[g(\pi, S_\alpha^\pi - 1) - g(\pi, S_\alpha^\pi)] + c((\widetilde{u}_1 - u_1) + \alpha(\widetilde{u}_2 - u_2)) \\ &= \alpha^l\left(b\mathbb{P}\left\{\sum_{n=1}^{l+1} w_n \geq S_\alpha^\pi \middle| \pi^1 = \pi\right\} - h\mathbb{P}\left\{\sum_{n=1}^{l+1} w_n \leq S_\alpha^\pi - 1 \middle| \pi^1 = \pi\right\}\right) - (1-\alpha)c \\ &= \alpha^l\left((b+h)\mathbb{P}\left\{\sum_{n=1}^{l+1} w_n \geq S_\alpha^\pi \middle| \pi^1 = \pi\right\} - h\right) - (1-\alpha)c. \end{aligned}$$

Since w_t is non-negative, $\forall t \in \mathbb{Z}_+$, Markov's inequality implies that $J_\alpha^{\widetilde{U}}(\pi, y) - J_\alpha^U(\pi, y) \leq \alpha^l\left((b+h)\frac{\mu(l+1)}{S_\alpha^\pi} - h\right) - (1-\alpha)c$. Since $S_\alpha^\pi > M$, $\alpha^l\left((b+h)\frac{\mu(l+1)}{S_\alpha^\pi} - h\right) - (1-\alpha)c < 0$. We have a contradiction because the expected total discounted cost under the alternative policy cannot be smaller than the expected total discounted cost under the optimal policy. We thus conclude that any policy with $S_\alpha^\pi > M$ for some $\pi \in \Pi$ cannot be optimal.

(ii) For any $\alpha \in (0,1)$, let $U = (u_1, u_2, \ldots)$ represent the order quantities under the optimal belief-dependent base-stock levels S_α^π, $\forall \pi \in \Pi$. Suppose that $\exists \pi \in \Pi$ such that $S_\alpha^\pi < 0$. Now consider all sample paths that start with $y_1 = y$ for some finite $y \leq S_\alpha^\pi$ and $\pi^1 = \pi$ where $S_\alpha^\pi < 0$. Let K be the first period after period 1 with a replenishment order, i.e., $K = \min_{n \in \mathbb{Z}_+} \{n : n \geq 2, u_n > 0 | \pi^1 = \pi, y_1 = y\}$. For a given sample path, if $K = k$, we construct an alternative policy with order quantities $\widetilde{U} = (\widetilde{u}_1, \widetilde{u}_2, \ldots)$ such that

$$\widetilde{u}_t = \begin{cases} u_1 + 1 & \text{if } t = 1, \\ u_k - 1 & \text{if } t = k, \\ u_t & \text{otherwise.} \end{cases}$$

The inventory position plus the order quantity in period t under the alternative policy is

$$\widetilde{y}_t + \widetilde{u}_t = \begin{cases} y_t + u_t + 1 & \text{if } 1 \leq t \leq k-1, \\ y_t + u_t & \text{if } t \geq k. \end{cases}$$

Note that $y_t + u_t < 0$ and $\widetilde{y}_t + \widetilde{u}_t \leq 0$ for $t < k$. Hence:

$$J_\alpha^{\widetilde{U}}(\pi, y) - J_\alpha^U(\pi, y) = \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1}[c\widetilde{u}_t + \alpha^l g(\pi^t, \widetilde{y}_t + \widetilde{u}_t) - cu_t - \alpha^l g(\pi^t, y_t + u_t)] \middle| \pi^1 = \pi, y_1 = \widetilde{y}_1 = y\right]$$

$$= \sum_{k=2}^{\infty} \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1}[c\widetilde{u}_t + \alpha^l g(\pi^t, \widetilde{y}_t + \widetilde{u}_t) - cu_t - \alpha^l g(\pi^t, y_t + u_t)] \middle| K = k, \pi^1 = \pi, y_1 = \widetilde{y}_1 = y\right] \mathbb{P}\{K = k\}$$

$$= \sum_{k=2}^{\infty} \left[c - \alpha^{k-1} c - \alpha^l b \sum_{t=1}^{k-1} \alpha^{t-1}(\widetilde{y}_t + \widetilde{u}_t - y_t - u_t)\right] \mathbb{P}\{K = k\}$$

$$= \sum_{k=2}^{\infty} [(1-\alpha)c - \alpha^l b] \left(\frac{1-\alpha^{k-1}}{1-\alpha}\right) \mathbb{P}\{K = k\}.$$

Since $\frac{1-\alpha^{k-1}}{1-\alpha} > 0$, $\mathbb{P}\{K = k\} > 0$ for some $k \geq 2$, and $(1-\alpha)c - \alpha^l b < 0$ (recall $\alpha^l b > c$), we have $J_\alpha^{\widetilde{U}}(\pi, y) - J_\alpha^U(\pi, y) < 0$. We have a contradiction because the expected total discounted cost under the alternative policy cannot be smaller than the expected total discounted cost under the optimal policy. Any policy with $S_\alpha^\pi < 0$ for some $\pi \in \Pi$ cannot be optimal. □

A similar threshold policy is available in the literature: Treharne and Sox (2002) establish the optimality of a belief-dependent base-stock policy for a finite-horizon total-cost problem with bounded Markov-modulated demand. We extend the optimal policy structure in Treharne and Sox (2002) to an *infinite-horizon discounted-cost* problem with *more general* Markov-modulated demand (not necessarily bounded). In Section 4.2, we further extend the optimal policy structure in Treharne and Sox (2002) to the infinite-horizon average-cost problem.

4.2. The Average-Cost Problem

We next consider the vanishing discount method for our analysis of the average-cost problem: For a fixed $\bar{\pi} \in \Pi$, we define $\delta_\alpha(\pi, y) := v_\alpha(\pi, y) - v_\alpha(\bar{\pi}, 0)$ as the differential discounted cost function, $\forall \pi \in \Pi$, $\forall y \in \mathbb{Z}$. For any $\alpha \in (0,1)$, the equation in (3) implies that

$$\delta_\alpha(\pi, y) + (1-\alpha) v_\alpha(\bar{\pi}, 0) = \min_{u \geq 0} \left\{ cu + \alpha^l g(\pi, y+u) + \alpha \sum_{w=0}^{\infty} \delta_\alpha(T(\pi, w), y+u-w) \hat{r}_\pi(w) \right\}. \tag{4}$$

We will show (in Theorem 1) that there exists a constant λ^* and a Lipschitz continuous function $\delta^*(\cdot,\cdot)$ that together satisfy the average-cost optimality equation:

$$\delta^*(\pi, y) + \lambda^* = \min_{u \geq 0} \left\{ cu + g(\pi, y+u) + \sum_{w=0}^{\infty} \delta^*(T(\pi, w), y+u-w) \hat{r}_\pi(w) \right\}, \ \forall \pi \in \Pi, \ \forall y \in \mathbb{Z},$$

such that $(1-\alpha) v_\alpha(\bar{\pi}, 0) \to \lambda^*$ and $\delta_\alpha(\pi, y) \to \delta^*(\pi, y)$ as α goes to 1. In order to obtain this analytical result, we establish that $(1-\alpha) v_\alpha(\bar{\pi}, 0)$ is bounded with respect to $\alpha \in (0,1)$ (see Lemma 1), and that $\delta_\alpha(\cdot,\cdot)$ is Lipschitz continuous and uniformly bounded with respect to $\alpha \in (0,1)$ (see Lemma 2). We will also show (in Theorem 1) that the optimal replenishment policy is a belief-dependent base-stock policy in our average-cost problem.

LEMMA 1. *$(1-\alpha) v_\alpha(\bar{\pi}, 0)$ is bounded with respect to $\alpha \in (0,1)$.*

Proof. For any $\alpha \in (0,1)$ and the initial inventory position $\widetilde{y}_1 = 0$, consider a replenishment policy with order quantities $\widetilde{U} = (\widetilde{u}_1, \widetilde{u}_2, \ldots)$ such that $\widetilde{u}_t = 0$ if $t = 1$ and $\widetilde{u}_t = w_{t-1}$ if $t > 1$. Note that the above policy implies that the base-stock level is zero for all $\pi \in \Pi$. Thus $\widetilde{y}_t + \widetilde{u}_t = 0$, $\forall t \in \mathbb{Z}_+$. Since this is a suboptimal policy, we have

$$\begin{aligned}
(1-\alpha) v_\alpha(\bar{\pi}, 0) &\leq (1-\alpha) J_\alpha^{\widetilde{U}}(\bar{\pi}, 0) \\
&= (1-\alpha) \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} [c\widetilde{u}_t + \alpha^l g(\pi^t, \widetilde{y}_t + \widetilde{u}_t)] \middle| \pi^1 = \bar{\pi}, \widetilde{y}_1 = 0 \right] \\
&= (1-\alpha) \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} \left(c\widetilde{u}_t + \alpha^l b \sum_{n=0}^{l} w_{t+n} \right) \middle| \pi^1 = \bar{\pi} \right] \\
&= (1-\alpha) \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} \left(\alpha c w_t + \alpha^l b \sum_{n=0}^{l} w_{t+n} \right) \middle| \pi^1 = \bar{\pi} \right] \\
&= (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1} \left(\alpha c \mathbb{E}[w_t | \pi^1 = \bar{\pi}] + \alpha^l b \sum_{n=0}^{l} \mathbb{E}[w_{t+n} | \pi^1 = \bar{\pi}] \right) \\
&\leq (1-\alpha) \left[\sum_{t=1}^{\infty} \alpha^{t-1} \right] [c + b(l+1)]\mu = [c + b(l+1)]\mu.
\end{aligned}$$

Hence $(1-\alpha) v_\alpha(\bar{\pi}, 0)$ is bounded with respect to $\alpha \in (0,1)$. □

In order to obtain further analytical results, we assume that the Markov chain governing the demand state transitions is ergodic. Previous work has required the *irreducibility* of the underlying

Markov chain for optimal policy characterization in average-cost inventory models with *perfectly* observed Markov-modulated demand (see Beyer and Sethi 1997 and Huh et al. 2011). In this study, in addition to irreducibility, we also require the *aperiodicity* of the underlying Markov chain for our average-cost inventory model with *partially* observed Markov-modulated demand.

Assumption 1. *The Markov chain with transition matrix P is ergodic.*

We now consider two demand state processes $\{d_t\}_{t\in\mathbb{Z}_+}$ and $\{\tilde{d}_t\}_{t\in\mathbb{Z}_+}$, both evolving according to Markov chains with the same transition matrix. Let $\nu(i,j) := \mathbb{P}\{d=i, \tilde{d}=j\}$ denote an arbitrary joint probability mass function for demand states d and $\tilde{d}$. Also, let $V_{\pi,\widetilde{\pi}} := \Big\{\nu : \sum_{j\in\mathcal{N}}\nu(i,j)=\pi_i, \forall i\in\mathcal{N}, \text{ and } \sum_{i\in\mathcal{N}}\nu(i,j)=\widetilde{\pi}_j, \forall j\in\mathcal{N}\Big\}$. Following Borkar (2000), we define the Wasserstein distance between two beliefs π and $\widetilde{\pi}$ that correspond to d and $\tilde{d}$, respectively:

$$\Delta(\pi,\widetilde{\pi}) := \inf_{\nu\in V_{\pi,\widetilde{\pi}}}\{\mathbb{E}_\nu[|d-\tilde{d}|]\} = \inf_{\nu\in V_{\pi,\widetilde{\pi}}}\left\{\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}|i-j|\nu(i,j)\right\}.$$

Using the above definition and our structural results in Proposition 1, under Assumption 1, Lemma 2 proves that $\delta_\alpha(\cdot,\cdot)$ is Lipschitz continuous and uniformly bounded with respect to $\alpha\in(0,1)$.

Lemma 2. *Suppose that Assumption 1 holds. For all finite $Y\in\mathbb{Z}_+$, $\delta_\alpha:\Pi\times\mathbb{Y}\to\mathbb{R}$, where $\mathbb{Y}:=\{-Y,-Y+1,\ldots,Y-1,Y\}$, is Lipschitz continuous and uniformly bounded with respect to $\alpha\in(0,1)$.*

Proof. We consider two systems with initial beliefs π and $\widetilde{\pi}$, and initial inventory positions $y\in\mathbb{Y}$ and $\widetilde{y}\in\mathbb{Y}$, respectively. Let $y_1,y_2,\ldots$ denote the inventory positions observed in the first system with beliefs $\pi^1,\pi^2,\ldots$ under the optimal belief-dependent base-stock policy with order quantities $U=(u_1,u_2,\ldots)$. Similarly, let $\widetilde{y}_1,\widetilde{y}_2,\ldots$ denote the inventory positions observed in the second system with beliefs $\widetilde{\pi}^1,\widetilde{\pi}^2,\ldots$ under an alternative policy with order quantities $\widetilde{U}=(\widetilde{u}_1,\widetilde{u}_2,\ldots)$ such that $\widetilde{u}_t=\max\{(y_t+u_t)-\widetilde{y}_t,0\},\ \forall t\in\mathbb{Z}_+$.

Let $\eta_{ij}:=\min\{n\in\mathbb{Z}_+: d_n=\tilde{d}_n|d_1=i,\tilde{d}_1=j\}$ denote the first period in which the demand states of these two systems become equal to each other, given that the initial demand state is i in the first system and j in the second system. Also, let $\tilde{K}_{ij}:=\min_{k\in\mathbb{Z}_+}\{k\geq n: y_k+u_k=\widetilde{y}_k+\widetilde{u}_k|\eta_{ij}=n\}$ denote the first period after period η_{ij} in which the inventory positions of these two systems become equal to each other. Following the coupling argument in Borkar (2000), we verify that the same demand values are observed in these systems with initial beliefs π and $\widetilde{\pi}$ once the demand states d_t and $\tilde{d}_t$ become equal to each other. Hence, if the inventory positions become equal to each other as well in a certain period, they will remain equal to each other in all future periods, i.e., if $\tilde{K}_{ij}=k$, then $y_t=\widetilde{y}_t$ and $u_t=\widetilde{u}_t$, $\forall t\geq k+1$. Thus:

$$\delta_\alpha(\widetilde{\pi},\widetilde{y})-\delta_\alpha(\pi,y)=v_\alpha(\widetilde{\pi},\widetilde{y})-v_\alpha(\bar{\pi},0)-v_\alpha(\pi,y)+v_\alpha(\bar{\pi},0)=v_\alpha(\widetilde{\pi},\widetilde{y})-v_\alpha(\pi,y)\leq J_\alpha^{\widetilde{U}}(\widetilde{\pi},\widetilde{y})-v_\alpha(\pi,y)$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{\infty}\alpha^{t-1}[c\widetilde{u}_t + \alpha^l g(\widetilde{\pi}^t, \widetilde{y}_t + \widetilde{u}_t) - cu_t - \alpha^l g(\pi^t, y_t + u_t)]\middle|\pi^1 = \pi, \widetilde{\pi}^1 = \widetilde{\pi}, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \\
&\leq \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\sum_{n=1}^{\infty}\sum_{k=n}^{\infty}\mathbb{E}\left[\sum_{t=1}^{k}\alpha^{t-1}[\alpha^l[g(\widetilde{\pi}^t, \widetilde{y}_t + \widetilde{u}_t) - g(\pi^t, y_t + u_t)] + c(\widetilde{u}_t - u_t)]\middle| d_1 = i, \tilde{d}_1 = j, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \\
&\qquad\qquad \mathbb{P}\{\tilde{K}_{ij} = k\}\mathbb{P}\{\eta_{ij} = n\}\pi_i\widetilde{\pi}_j. \qquad (5)
\end{aligned}
$$

By the alternative policy structure, $y_t + u_t \leq \widetilde{y}_t + \widetilde{u}_t$, $\forall t \in \mathbb{Z}_+$. By Proposition 1, $0 \leq y_t + u_t \leq \widetilde{y}_t + \widetilde{u}_t \leq \max\{M, Y\}$, $\forall t \in \mathbb{Z}_+$. Hence:

$$
\begin{aligned}
g(\widetilde{\pi}^t, \widetilde{y}_t + \widetilde{u}_t) &= h\mathbb{E}\left[\widetilde{y}_t + \widetilde{u}_t - \sum_{m=0}^{l}\widetilde{w}_{t+m}\middle|\sum_{m=0}^{l}\widetilde{w}_{t+m} \leq \widetilde{y}_t + \widetilde{u}_t, \tilde{d}_1 = j\right]\mathbb{P}\left\{\sum_{m=0}^{l}\widetilde{w}_{t+m} \leq \widetilde{y}_t + \widetilde{u}_t\middle|\tilde{d}_1 = j\right\} \\
&\quad + b\mathbb{E}\left[\sum_{m=0}^{l}\widetilde{w}_{t+m} - \widetilde{y}_t - \widetilde{u}_t\middle|\sum_{m=0}^{l}\widetilde{w}_{t+m} > \widetilde{y}_t + \widetilde{u}_t, \tilde{d}_1 = j\right]\mathbb{P}\left\{\sum_{m=0}^{l}\widetilde{w}_{t+m} > \widetilde{y}_t + \widetilde{u}_t\middle|\tilde{d}_1 = j\right\} \\
&\quad \leq h\max\{M, Y\} + b(l+1)\mu = A.
\end{aligned}
$$

Note that $A \in \mathbb{R}_+$ is finite. Consequently:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{k}\alpha^{t+l-1}[g(\widetilde{\pi}^t, \widetilde{y}_t + \widetilde{u}_t) - g(\pi^t, y_t + u_t)]\middle| d_1 = i, \tilde{d}_1 = j, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{k-1} g(\widetilde{\pi}^t, \widetilde{y}_t + \widetilde{u}_t)\middle| d_1 = i, \tilde{d}_1 = j, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \leq (k-1)A. \qquad (6)
\end{aligned}
$$

Recall that $y_{k+1} = \widetilde{y}_{k+1}$ and $w_t = \widetilde{w}_t$, $\forall t \geq n$. Also, recall that $\mathbb{E}[w_t] \leq \mu$ and $\mathbb{E}[\widetilde{w}_t] \leq \mu$, $\forall t \in \mathbb{Z}_+$. Thus:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{k}\alpha^{t-1}c(\widetilde{u}_t - u_t)\middle| d_1 = i, \tilde{d}_1 = j, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{k}c(\widetilde{y}_{t+1} - \widetilde{y}_t + \widetilde{w}_t - y_{t+1} + y_t - w_t)\middle| d_1 = i, \tilde{d}_1 = j, y_1 = y, \widetilde{y}_1 = \widetilde{y}\right] \\
&= c(y - \widetilde{y}) + c\sum_{t=1}^{n-1}(\mathbb{E}[\widetilde{w}_t|\tilde{d}_1 = j] - \mathbb{E}[w_t|d_1 = i]) \leq c(y - \widetilde{y}) + c(n-1)\mu. \qquad (7)
\end{aligned}
$$

The inequalities in (5)–(7) imply the following inequalities.

$$
\begin{aligned}
\delta_\alpha(\widetilde{\pi}, \widetilde{y}) - \delta_\alpha(\pi, y) &\leq \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\sum_{n=1}^{\infty}\sum_{k=n}^{\infty}[(k-1)A + c(y - \widetilde{y}) + c(n-1)\mu]\mathbb{P}\{\tilde{K}_{ij} = k\}\mathbb{P}\{\eta_{ij} = n\}\pi_i\widetilde{\pi}_j \\
&\leq A\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\mathbb{E}[\tilde{K}_{ij} - 1]\pi_i\widetilde{\pi}_j + c\mu\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}\mathbb{E}[\eta_{ij} - 1]\pi_i\widetilde{\pi}_j + c(y - \widetilde{y}). \qquad (8)
\end{aligned}
$$

We make the following two observations regarding the inequality in (8):

(1) Under Assumption 1, there exists a finite $\Gamma > 0$ such that $\mathbb{E}[\eta_{ij} - 1] \leq \Gamma$, $\forall i, j \in \mathcal{N}$ (see Borkar 2000). If $i = j$, $\mathbb{E}[\eta_{ij}] = 1$. If $i \neq j$, $|i - j| \geq 1$. Thus $\mathbb{E}[\eta_{ij} - 1] \leq \Gamma|i - j|$, $\forall i, j \in \mathcal{N}$.

(2) If $i = j$, the inventory positions become equal once a replenishment order is observed in the second system. Since $S_\alpha^\pi \geq 0$, $\forall \pi \in \Pi$, by Proposition 1, we place an order in the second system no later than the period up to which a total of $\widetilde{y} + 1$ units of demand are observed. Hence,

$\tilde{K}_{ij} \le \min\{n : \sum_{t=1}^{n} \widetilde{w}_t \ge \widetilde{y}+1 |\ \widetilde{\pi}^1 = \widetilde{\pi}\}$. For a sample path starting with belief $\hat{\pi} \in \Pi$, and for a finite $\xi \in \mathbb{Z}_+$, let $\tau_{\hat{\pi},\xi} := \min\{n : \sum_{t=1}^{n} \widetilde{w}_t \ge \xi |\ \widetilde{\pi}^1 = \hat{\pi}\}$ be the first period when the cumulative demand is no less than ξ. Using the second moment method, we show that

$$\mathbb{P}\{\widetilde{w}_t \ge 1 | \widetilde{\pi}^t\} \ge \frac{(\mathbb{E}[\widetilde{w}_t | \widetilde{\pi}^t])^2}{\mathbb{E}[\widetilde{w}_t^2 | \widetilde{\pi}^t]} = \frac{\mu^2_{\widetilde{\pi}^t}}{\zeta_{\widetilde{\pi}^t}} = \frac{\left(\sum_{i\in\mathcal{N}} \mu_i \widetilde{\pi}_i^t\right)^2}{\sum_{i\in\mathcal{N}} \zeta_i \widetilde{\pi}_i^t}.$$

As we assume $\exists i \in \mathcal{N}$ such that $\mu_i > 0$, and by Assumption 1, we have $\mathbb{P}\{\widetilde{w}_t \ge 1 | \widetilde{\pi}^t\} > 0$, $\forall t \in \mathbb{Z}_+$. Thus $\mathbb{P}\left\{\sum_{t=1}^{\xi} \widetilde{w}_t \ge \xi \middle| \widetilde{\pi}^1 = \hat{\pi}\right\} > 0$, $\forall \hat{\pi} \in \Pi$. Let $\rho_\xi := \max_{\hat{\pi}\in\Pi} \left\{\mathbb{P}\left[\sum_{t=1}^{\xi} \widetilde{w}_t < \xi \middle| \widetilde{\pi}^1 = \hat{\pi}\right]\right\} < 1$. Notice that

$$\begin{aligned}
\mathbb{E}[\tau_{\hat{\pi},\xi}] &= \sum_{n=0}^{\xi-1} \mathbb{P}\{\tau_{\hat{\pi},\xi} > n\} + \sum_{n=\xi}^{\infty} \mathbb{P}\{\tau_{\hat{\pi},\xi} > n\} \le \xi + \sum_{n=\xi}^{\infty} \mathbb{P}\{\tau_{\hat{\pi},\xi} > n\} \\
&= \xi + \sum_{n=\xi}^{\infty} \mathbb{P}\left\{\sum_{t=1}^{n} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} = \xi + \sum_{k=1}^{\infty} \sum_{m=k\xi}^{(k+1)\xi-1} \mathbb{P}\left\{\sum_{t=1}^{m} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} \\
&\le \xi + \sum_{k=1}^{\infty} \sum_{m=k\xi}^{(k+1)\xi-1} \mathbb{P}\left\{\sum_{t=1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} = \xi + \sum_{k=1}^{\infty} \xi \ \mathbb{P}\left\{\sum_{t=1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} \\
&\le \xi + \xi \sum_{k=1}^{\infty} \mathbb{P}\left\{\sum_{t=1}^{\xi} \widetilde{w}_t < \xi, \sum_{t=\xi+1}^{2\xi} \widetilde{w}_t < \xi, \ldots, \sum_{t=(k-1)\xi+1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\}.
\end{aligned}$$

Also notice that

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{t=1}^{\xi} \widetilde{w}_t < \xi, \sum_{t=\xi+1}^{2\xi} \widetilde{w}_t < \xi, \ldots, \sum_{t=(k-1)\xi+1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} \\
&= \mathbb{P}\left\{\sum_{t=\xi+1}^{2\xi} \widetilde{w}_t < \xi, \ldots, \sum_{t=(k-1)\xi+1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}, \sum_{t=1}^{\xi} \widetilde{w}_t < \xi\right\} \mathbb{P}\left\{\sum_{t=1}^{\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} \\
&\le \mathbb{P}\left\{\sum_{t=\xi+1}^{2\xi} \widetilde{w}_t < \xi, \ldots, \sum_{t=(k-1)\xi+1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^{\xi+1} = \breve{\pi}\right\} \rho_\xi
\end{aligned}$$

for some $\breve{\pi} \in \Pi$. Proceeding similarly, it can be shown that

$$\mathbb{P}\left\{\sum_{t=1}^{\xi} \widetilde{w}_t < \xi, \sum_{t=\xi+1}^{2\xi} \widetilde{w}_t < \xi, \ldots, \sum_{t=(k-1)\xi+1}^{k\xi} \widetilde{w}_t < \xi \ \middle|\ \widetilde{\pi}^1 = \hat{\pi}\right\} \le \rho_\xi^k.$$

Consequently:

$$\mathbb{E}[\tau_{\hat{\pi},\xi}] \le \xi + \xi \sum_{k=1}^{\infty} \rho_\xi^k = \frac{\xi}{1-\rho_\xi} < \infty.$$

Thus if $i = j$, because $\tilde{K}_{ij} \le \tau_{\widetilde{\pi},\widetilde{y}+1}$, we obtain $\mathbb{E}[\tilde{K}_{ij}] < \infty$. If $i \ne j$, because $\tilde{K}_{ij} \le \eta_{ij} + \tau_{\hat{\pi},\hat{y}+1}$ for some $\hat{\pi} \in \Pi$ and $\hat{y} \in \mathbb{Y}$, and $\mathbb{E}[\eta_{ij}] < \infty$, we again obtain $\mathbb{E}[\tilde{K}_{ij}] < \infty$. Hence there exists a finite $B \in \mathbb{R}_+$ such that $\mathbb{E}[\tilde{K}_{ij} - 1] \le B(|i-j| + |y - \widetilde{y}|)$.

Now recall the inequality in (8):

$$\delta_\alpha(\widetilde{\pi}, \widetilde{y}) - \delta_\alpha(\pi, y) \le A \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} \mathbb{E}[\tilde{K}_{ij} - 1]\pi_i \widetilde{\pi}_j + c\mu \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} \mathbb{E}[\eta_{ij} - 1]\pi_i \widetilde{\pi}_j + c(y - \widetilde{y})$$

$$\begin{aligned}
&\leq A\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} B(|i-j|+|y-\widetilde{y}|)\pi_i\widetilde{\pi}_j + c\mu\Gamma\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}|i-j|\pi_i\widetilde{\pi}_j + c|y-\widetilde{y}| \\
&= (AB+c\mu\Gamma)\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}}|i-j|\pi_i\widetilde{\pi}_j + (AB+c)|y-\widetilde{y}| \\
&= (AB+c\mu\Gamma)\,\mathbb{E}[|d_1-\tilde{d}_1|] + (AB+c)|y-\widetilde{y}|. \qquad (9)
\end{aligned}$$

Pick an arbitrary $\varepsilon > 0$. With an appropriate choice of the joint mass function of $(d_1, \tilde{d}_1)$, we can obtain

$$\delta_\alpha(\widetilde{\pi},\widetilde{y}) - \delta_\alpha(\pi,y) \leq (AB+c\mu\Gamma)\left[\Delta(\pi,\widetilde{\pi})+\varepsilon\right] + (AB+c)|y-\widetilde{y}|.$$

Since $\varepsilon > 0$ is arbitrary, there exists a finite $C\in\mathbb{R}_+$ such that $\delta_\alpha(\widetilde{\pi},\widetilde{y}) - \delta_\alpha(\pi,y) \leq C[\Delta(\pi,\widetilde{\pi}) + |y-\widetilde{y}|]$. Since $\pi,\widetilde{\pi}\in\Pi$ and $y,\widetilde{y}\in\mathbb{Z}$ are arbitrary, $|\delta_\alpha(\widetilde{\pi},\widetilde{y}) - \delta_\alpha(\pi,y)| \leq C[\Delta(\pi,\widetilde{\pi}) + |y-\widetilde{y}|]$. Thus $\delta_\alpha(\cdot,\cdot)$ is Lipschitz continuous for $\alpha\in(0,1)$.

Since the inequality in (9) holds for any $\pi,\widetilde{\pi}\in\Pi$ and for any $y,\widetilde{y}\in\mathbb{Z}$, and $\delta_\alpha(\bar{\pi},0) = v_\alpha(\bar{\pi},0) - v_\alpha(\bar{\pi},0) = 0$, the following inequality holds.

$$|\delta_\alpha(\pi,y)| = |\delta_\alpha(\pi,y) - \delta_\alpha(\bar{\pi},0)| \leq (AB+c\mu\Gamma)\mathbb{E}[|d_1-\tilde{d}_1|] + (AB+c)|y|. \qquad (10)$$

Since $|d_1-\tilde{d}_1| \leq N$, there exists a finite $D\in\mathbb{R}_+$ such that $|\delta_\alpha(\pi,y)|\leq D$ for any $\pi\in\Pi$ and for any finite $y\in\mathbb{Z}$. Thus $\delta_\alpha(\cdot,\cdot)$ is uniformly bounded with respect to $\alpha\in(0,1)$. □

We are now ready to state the main result of this paper that builds upon Lemmas 1 and 2:

Theorem 1. *Under Assumption 1, there exist a constant λ^* and a Lipschitz continuous function $\delta^*(\cdot,\cdot)$ satisfying the following average-cost optimality equation*

$$\delta(\pi,y)+\lambda = \min_{u\geq 0}\left\{cu + g(\pi,y+u) + \sum_{w=0}^{\infty}\delta(T(\pi,w),y+u-w)\hat{r}_\pi(w)\right\}. \qquad (11)$$

Furthermore, there exists an optimal stationary inventory replenishment policy that can be described as a belief-dependent base-stock policy with base-stock levels $S^\pi, \forall\pi\in\Pi$.

Proof. By Lemma 2, $\delta_\alpha(\cdot,\cdot)$ is Lipschitz continuous and uniformly bounded with respect to $\alpha\in(0,1)$. By the Arzela-Ascoli Theorem, there exist a sequence $\alpha_t\to 1$ as $t\to\infty$ and a Lipschitz continuous function $\delta^*(\pi,y)$ such that $\delta_{\alpha_t}(\pi,y)\to\delta^*(\pi,y)$, $\forall\pi\in\Pi$ and for any finite $y\in\mathbb{Z}$. By Lemma 1, $(1-\alpha)v_\alpha(\bar{\pi},0)$ is bounded with respect to $\alpha\in(0,1)$. By the Bolzano-Weierstrass Theorem, there exist a subsequence $\alpha_{t_n}\to 1$ as $n\to\infty$ and a constant λ^* such that $(1-\alpha_{t_n})v_{\alpha_{t_n}}(\bar{\pi},0)\to\lambda^*$. We now define the random variable $\Psi_t(\pi,z) := \delta_{\alpha_t}(T(\pi,w),z-w)$. Thus $\mathbb{E}[\Psi_t(\pi,z)] = \sum_{w=0}^{\infty}\delta_{\alpha_t}(T(\pi,w),z-w)\hat{r}_\pi(w)$. Since $\mathbb{P}\{w_t\geq 1|\pi^t\}>0$, $\forall t\in\mathbb{Z}_+$, by Proposition 1, the inventory position takes values no greater than M after a sufficiently large number of periods. Hence, by the inequality in (10), the following inequalities hold.

$$\Psi_t(\pi,z) \leq (AB+c\mu\Gamma)N + (AB+c)|z-w| \leq (AB+c\mu\Gamma)N + (AB+c)(M+w).$$

Since $\mathbb{E}[w]$ is finite, w is integrable. Thus, by Lemma 2 and the dominated convergence theorem, $\Psi_t(\pi,z)$ converges everywhere to $\Psi(\pi,z) := \delta^*(T(\pi,w), z-w)$. By the inequality in (10), and since $\mathbb{E}[w]$ is finite, the following inequality holds.

$$\mathbb{E}[|\Psi_t(\pi,z)|] = \sum_{w=0}^{\infty} |\delta_{\alpha_t}(T(\pi,w), z-w)|\hat{r}_\pi(w) \leq (AB + c\mu\Gamma)N + (AB+c)\sum_{w=0}^{\infty}|w-z|\hat{r}_\pi(w) < \infty.$$

Hence $\Psi_t(\pi,z)$ is uniformly integrable with respect to t. Therefore, and since $\Psi_t(\pi,z)$ converges everywhere to $\Psi(\pi,z)$ for any fixed π and z, $\mathbb{E}[\Psi_t(\pi,z)] \to \mathbb{E}[\Psi(\pi,z)]$. This implies that $\lim_{\alpha_t\to 1}\sum_{w=0}^{\infty}\delta_{\alpha_t}(T(\pi,w), y+u-w)\hat{r}_\pi(w) = \sum_{w=0}^{\infty}\delta^*(T(\pi,w), y+u-w)\hat{r}_\pi(w)$. Consequently, using the equation in (4) and the bounded convergence theorem, we obtain

$$\delta^*(\pi,y) + \lambda^* = \min_{u\geq 0}\left\{cu + g(\pi, y+u) + \sum_{w=0}^{\infty}\delta^*(T(\pi,w), y+u-w)\hat{r}_\pi(w)\right\}. \tag{12}$$

Following Theorem 1 in Ross (1968), we verify the existence of an optimal stationary deterministic policy. By definition of $\delta_\alpha(\cdot,\cdot)$ and Proposition 1, $\delta_{\alpha_t}(\pi,y)$ is discrete-convex in y, $\forall \pi \in \Pi$. Since the limit of a sequence of discrete-convex functions is discrete-convex, $\delta^*(\pi,y)$ is also discrete-convex in y, $\forall \pi\in\Pi$. Thus the optimal stationary policy is a belief-dependent base-stock policy. □

In the literature several papers identify the optimal policy structure for average-cost inventory systems with Markov-modulated demand when the state of the underlying Markov chain is *perfectly* observed (Beyer and Sethi 1997 and Huh et al. 2011). To our knowledge, however, we are the first to characterize the optimal policy structure for average-cost inventory systems with Markov-modulated demand when the state of the underlying Markov chain can only be *partially* observed.

5. Discretized Approximation

Solving the optimality equation in (11) for each state $(\pi,y)\in\Pi\times\mathbb{Z}$ and finding the optimal base-stock level for each belief $\pi\in\Pi$ is a computational challenge since Π is an *uncountable* space. To address this challenge, we now discretize the uncountable space Π, on which the beliefs are defined, based on the regular grid approach developed by Lovejoy (1991): Let Q_n denote a regular grid for a given $n\in\mathbb{Z}_+$ such that the convex hull of Q_n is Π. Specifically, Q_n is defined by

$$Q_n := \left\{[\theta_1,...,\theta_N]\in\mathbb{Q}^N \;\middle|\; \theta_i = \frac{k_i}{n}, \sum_{i=1}^{N}k_i = n, k_i\in\mathbb{Z}_+\cup\{0\}\right\},$$

where $\mathbb{Q}$ is the set of rational numbers. (Recall that N is the number of demand states.) The number of grid points in Q_n is $\kappa_n = |Q_n| = \frac{(N-1+n)!}{(N-1)!n!}$. We denote the elements of Q_n by $\{q^1,\ldots,q^{\kappa_n}\}$. Any belief $\pi\in\Pi$ can be expressed as a convex combination of the grid points in Q_n, i.e., $\pi = \sum_{i=1}^{\kappa_n}\gamma_i(\pi)q^i$ where $\gamma_i(\pi)\geq 0$ denotes the convex combination multiplier associated with q^i, $\forall i = 1,\ldots,\kappa_n$, such that $\sum_{i=1}^{\kappa_n}\gamma_i(\pi) = 1$. Using the triangulation method developed by Lovejoy (1991), we find the smallest simplex containing the belief π and denote the corresponding convex combination multipliers by $\gamma_i^*(\pi)$. This method yields the convex representation scheme $\underset{\sim}{\gamma}_n := (\gamma_1^*(\cdot), \ldots, \gamma_{\kappa_n}^*(\cdot))$ with $\gamma_i^*(\cdot) > 0$ for at most N elements of Q_n as stated by the Carathéodory's Theorem.

Following Yu and Bertsekas (2004), we define ϵ_n as the fineness of the discretization scheme $(Q_n, \underset{\sim}{\gamma}_1)$ that is formulated as

$$\epsilon_n := \max_{\pi\in\Pi} \max_{q^i \in Q_n : \gamma_i^*(\pi) > 0} ||\pi - q^i||$$

where $||\cdot||$ denotes the Euclidean distance. Since Q_n is a regular grid and any belief can only be represented by the closest grid points to that belief according to our construction of $\underset{\sim}{\gamma}_n$, it can be shown that $\epsilon_n = \frac{\sqrt{N-1}}{n\sqrt{N}}$. Note that $\epsilon_n \to 0$ as $n \to \infty$. For any $n \in \mathbb{Z}_+$, we can compute the optimal average cost λ_n^* and the optimal differential cost function $\delta_n^*(\cdot,\cdot)$ associated with an ϵ_n-discretization scheme $(Q_n, \underset{\sim}{\gamma}_n)$ by solving the following optimality equations:

$$\delta_n(q,y) + \lambda_n = \min_{u\geq 0} \left\{ cu + g(q, y+u) + \sum_{w=0}^{\infty} \sum_{i=1}^{\kappa_n} \gamma_i^*(T(q,w))\delta_n(q^i, y+u-w)\hat{r}_q(w) \right\}, \ \forall q \in Q_n, \ \forall y \in \mathbb{Z}.$$

Numerical experiments in Section 6 verify the statement in Yu and Bertsekas (2004) that λ_n^* converges from below to the optimal average cost λ^* as n grows large. In addition, again following Yu and Bertsekas (2004), we derive an upper bound on the optimal average cost λ^*:

$$\lambda^* \leq \breve{\lambda}_n := \max_{(\pi,y)\in\Pi\times\mathbb{Z}} \left\{ \bar{\delta}_n(\pi,y) - \hat{\delta}_n(\pi,y) \right\} \tag{13}$$

where

$$\hat{\delta}_n(\pi,y) + \lambda_n^* = \min_{u\geq 0} \left\{ cu + g(\pi, y+u) + \sum_{w=0}^{\infty} \sum_{i=1}^{\kappa_n} \gamma_i^*(T(\pi,w))\delta_n^*(q^i, y+u-w)\hat{r}_\pi(w) \right\}, \ \forall \pi \in \Pi, \ \forall y \in \mathbb{Z},$$

and

$$\bar{\delta}_n(\pi,y) + \lambda_n^* = \min_{u\geq 0} \left\{ cu + g(\pi, y+u) + \sum_{w=0}^{\infty} \hat{\delta}_n(T(\pi,w), y+u-w)\hat{r}_\pi(w) \right\}, \ \forall \pi \in \Pi, \ \forall y \in \mathbb{Z}.$$

With the above discretization scheme, if the demand is bounded, we are able to address the computational challenge in our problem. Suppose that the demand is bounded above by W. We know from Proposition 1 that the optimal base-stock levels are between 0 and M in the discounted-cost problem. Following the same proof steps as in Proposition 1, these bounds can be shown to apply to the optimal base-stock levels in the average-cost problem, restricting the inventory position to take values between $-W$ and M. If the initial inventory position is above M or below $-W$, it eventually falls into this range after a finite number of periods. The contribution of the cost due to the excess or insufficient inventory in those initial periods to the average cost can thus be disregarded in our infinite-horizon planning. In Section 6, we use the lower bound λ_n^* obtained from the above discretization scheme in our numerical experiments on instances with bounded demand.

6. Numerical Results

In this section, for our MDP in Section 3, we conduct numerical experiments to investigate the value of implementing belief-dependent base-stock levels (see Section 6.1) and the performance of a myopic belief-dependent base-stock policy as a heuristic replenishment policy (see Section 6.2). We consider instances with three demand states, i.e., $\mathcal{N}$=$\{1,\ 2,\ 3\}$. The demand distributions are *Binomial*$(20,\ p)$, *Binomial*$(20,\ 0.5)$, and *Binomial*$(20,\ 1-p)$ for the demand states 1, 2, and 3, respectively. We generate a total of 108 instances in which $b=20$, $h\in\{2,5,10\}$, $l\in\{0,1,2\}$, $p\in\{0.1,0.2,0.3,0.4\}$, and the transition matrix P is

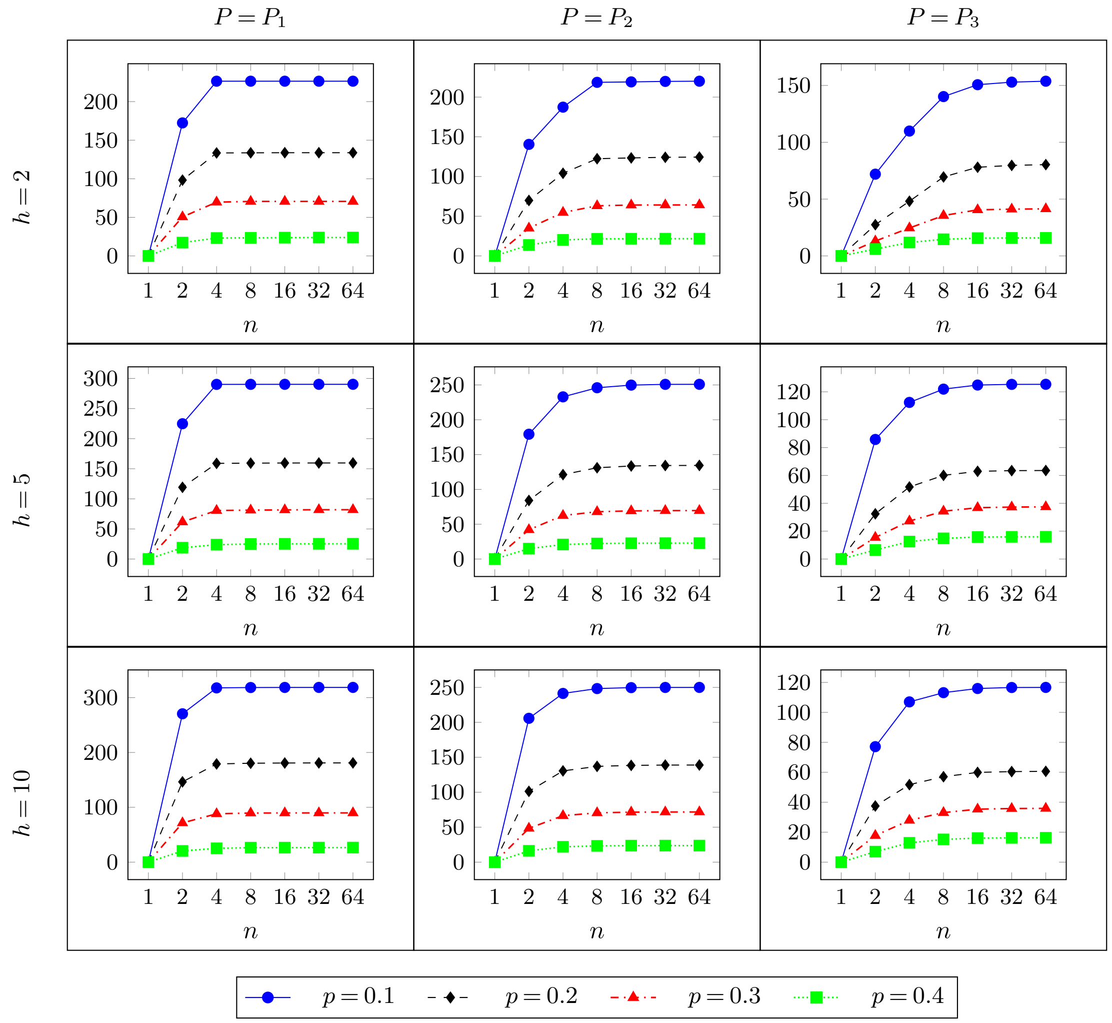
$$P_1=\begin{bmatrix}0.5 & 0.25 & 0.25\\ 0.25 & 0.5 & 0.25\\ 0.25 & 0.25 & 0.5\end{bmatrix},\ P_2=\begin{bmatrix}0.7 & 0.15 & 0.15\\ 0.15 & 0.7 & 0.15\\ 0.15 & 0.15 & 0.7\end{bmatrix},\ \text{or } P_3=\begin{bmatrix}0.9 & 0.05 & 0.05\\ 0.05 & 0.9 & 0.05\\ 0.05 & 0.05 & 0.9\end{bmatrix}.$$

Since the long-run average order quantity per period equals the expected demand per period under any base-stock policy, the unit ordering cost has no impact on the base-stock level calculation in our experiments, and thus we exclude it from our numerical study. Our instances are similar to those studied by many others (e.g., Chen 2010, Arifoğlu and Özekici 2010, 2011, Natarajan and Swaminathan 2014, and Chao et al. 2018) in terms of the lead time values and the shortage to holding cost ratios. Notice that Assumption 1 holds for each of our instances.

For each instance, using the policy iteration method, we calculate the average costs λ_n^* associated with our discretization scheme for $n\in\{1,2,4,8,16,32,64\}$ and their percentage differences from the average cost λ_1^*, i.e, $100\times\frac{\lambda_n^*-\lambda_1^*}{\lambda_1^*}$. Figures 1–3 exhibit these percentage gaps for our instances with lead times 0, 1, and 2, respectively. Notice that λ_1^* is the worst lower bound that can be obtained from our discretization scheme. It is important to note that, when $n=1$ in our discretization scheme, the number of grid points in Q_1 equals the number of demand states, i.e., $\kappa_1=N$. And there is a bijective relationship between the grid points and the demand states so that each grid point corresponds to a different demand state. Hence, solving the optimality equations associated with the ϵ_1-discretization scheme $(Q_1,\underline{\gamma}_1)$, we obtain the optimal average cost that could be achieved if the demand states were *perfectly* observed, which is λ_1^*, and an optimal policy that can be characterized as a state-dependent base-stock policy (Beyer and Sethi 1997). We observe from Figures 1–3 that the average costs λ_{64}^* and λ_{32}^* are virtually equal for each instance. Thus, we take the average cost λ_{64}^* as the optimal average cost in our experiments. We note that the difference between λ_{64}^* and λ_1^* can be viewed as the value of perfect state information.

Our simulation studies in Sections 6.1–6.2 consist of 30 replications of 10000 periods in each instance. We observed that the 10000-period horizon is long enough to represent our infinite-horizon

Figure 1 $100 \times \frac{\lambda_n^* - \lambda_1^*}{\lambda_1^*}$ vs. n when $l = 0$, $b = 20$, $P \in \{P_1, P_2, P_3\}$, $p \in \{0.1, 0.2, 0.3, 0.4\}$, $h \in \{2, 5, 10\}$.

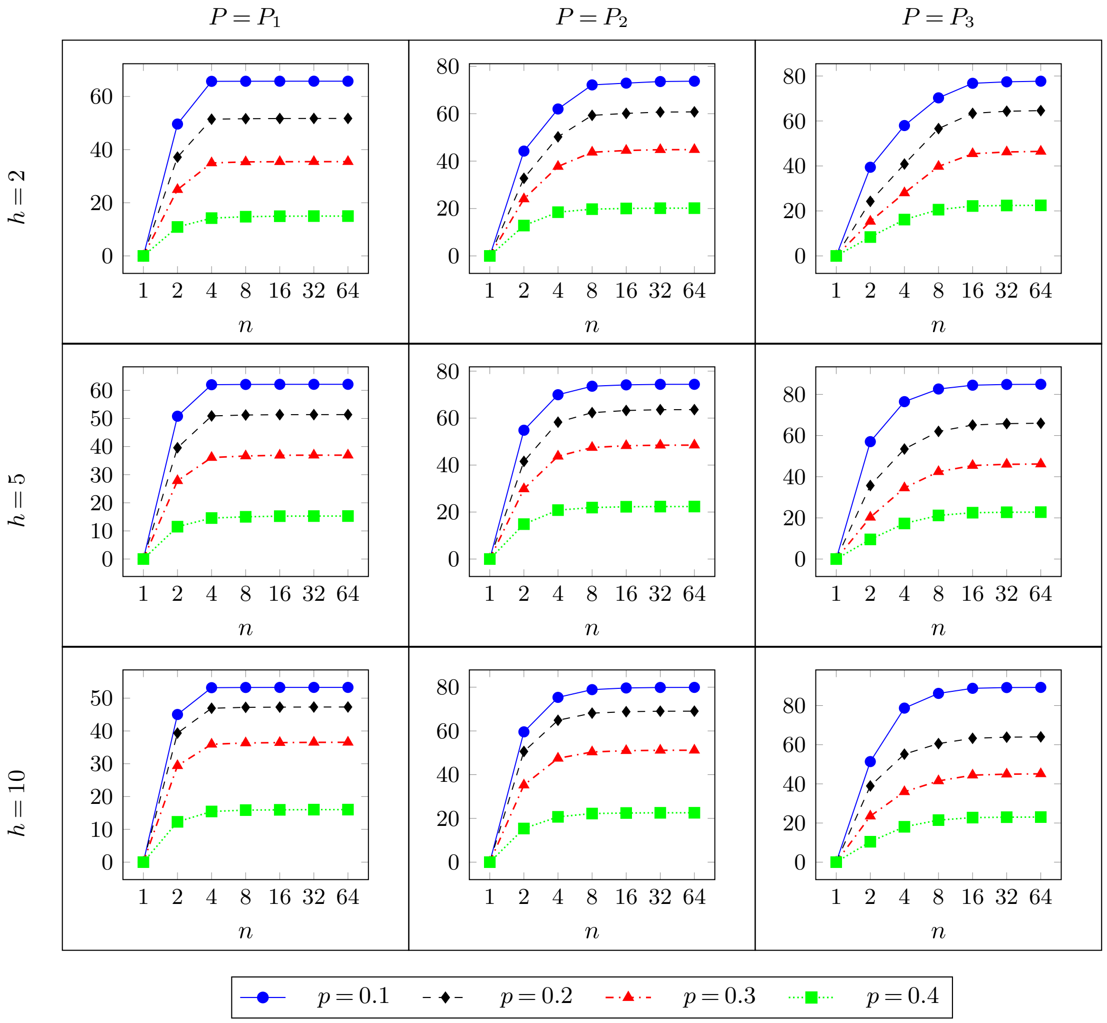


planning, and that the average cost found via simulation is not affected by the initial system state in our instances. We set the initial inventory position to be zero, while we randomly choose the initial state beliefs. We take into account the holding and backordering costs incurred in periods $l+1$ through 10000, excluding those incurred in periods 1 through l, in our average cost calculation. Starting from period 2, we update the state belief at the beginning of each period.

6.1. The Value of Belief-Dependent Base-Stock Levels

In order to investigate the value of Bayesian updating along with belief-dependent base-stock levels in our MDP in Section 3, first, we consider a much simpler MDP with a stationary demand distribution that is obtained by compounding the demand distributions based on the stationary distribution of the underlying Markov chain. For such an MDP, a myopic base-stock policy with a single base-stock level is optimal (Veinott 1965) and the optimal base-stock level can be eas-

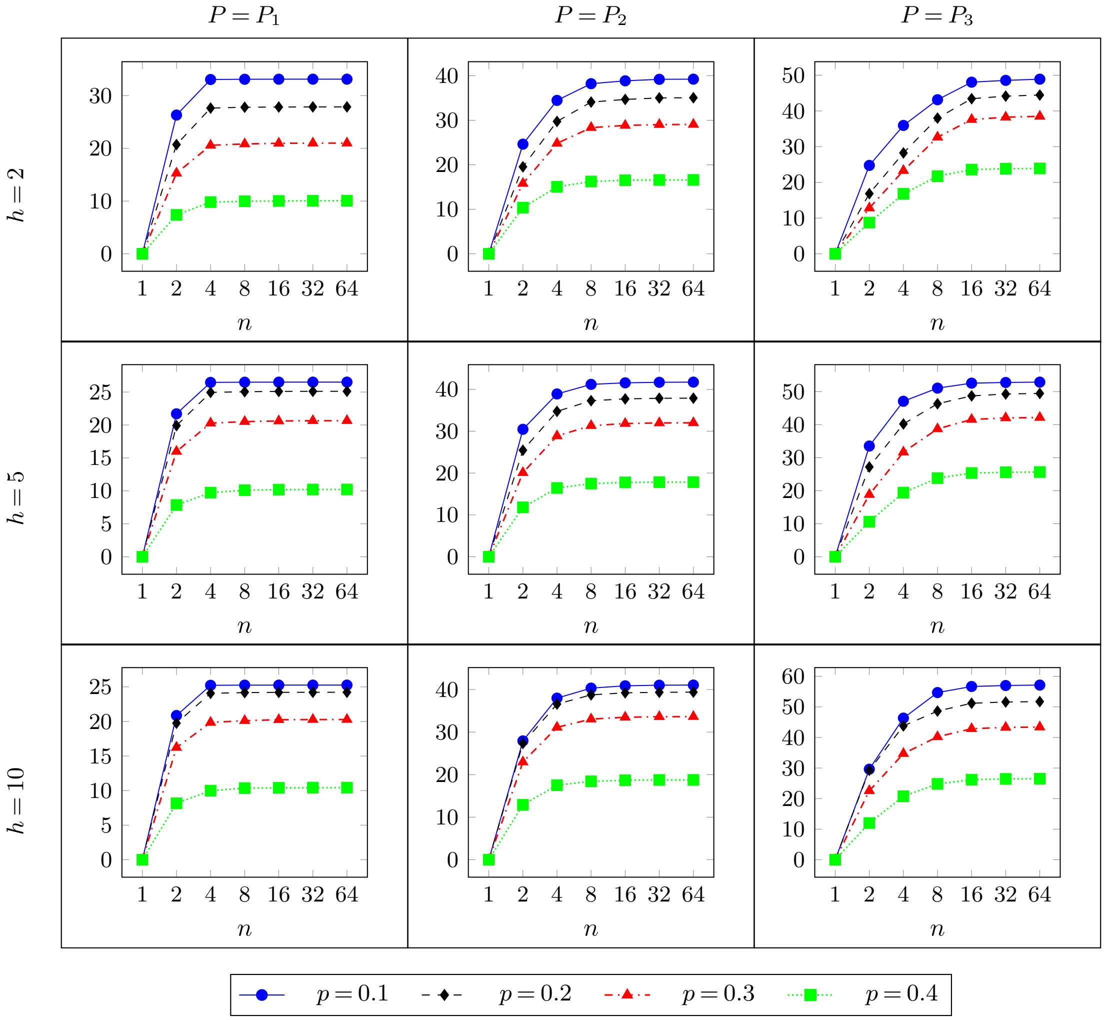
Figure 2 $100 \times \frac{\lambda_n^* - \lambda_1^*}{\lambda_1^*}$ vs. n when $l = 1$, $b = 20$, $P \in \{P_1, P_2, P_3\}$, $p \in \{0.1, 0.2, 0.3, 0.4\}$, $h \in \{2, 5, 10\}$.



ily found using the newsvendor formula applied to the lead-time demand distribution. For each instance, we calculate the optimal base-stock level for this MDP and simulate the inventory system under this base-stock level. Our simulation results indicate the average cost under this base-stock level is on average 16.1% greater than the optimal average cost λ_{64}^* on our test bed, highlighting the importance of incorporating the non-stationarity of demand distribution and the partial information about demand distribution into decision-making via Bayesian updating along with belief-dependent base-stock levels.

Recall that the optimal policy under perfectly observed demand states can be specified as a state-dependent base-stock policy. With this observation, we now consider another simplification of our MDP in Section 3 that in each period estimates the demand state and uses the corresponding base-stock level that is obtained from the ϵ_1-discretization scheme $(Q_1, \underline{\gamma}_1)$. As our estimate of the demand state in each period t, following Chapter 9 in Barber (2012), we choose the state

Figure 3 $100 \times \frac{\lambda_n^* - \lambda_1^*}{\lambda_1^*}$ vs. n when $l = 2,\ b = 20,\ P \in \{P_1, P_2, P_3\},\ p \in \{0.1, 0.2, 0.3, 0.4\},\ h \in \{2, 5, 10\}$.



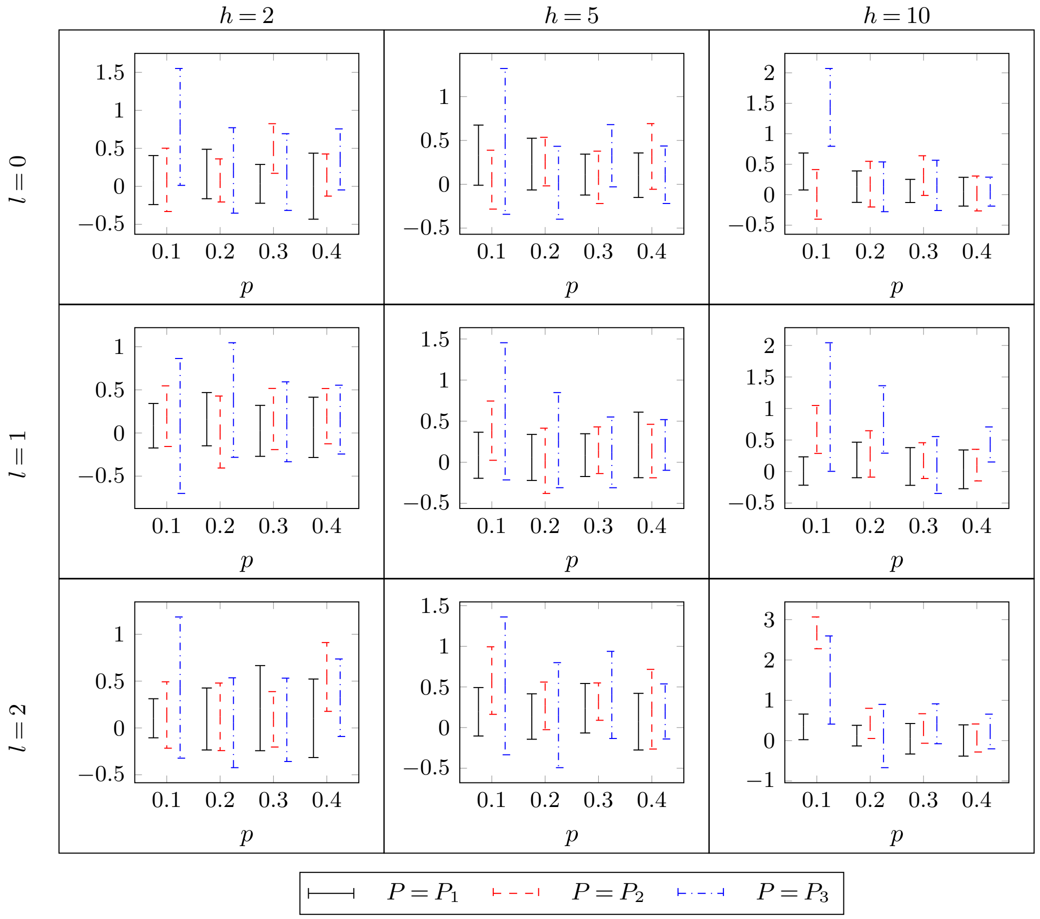
with the highest posterior probability based on the entire demand history, i.e., $\arg\max_{i \in \mathcal{N}} \pi_i^t$. Our simulation results show that the average cost under this simplification is on average 9.53% greater than the optimal average cost λ_{64}^* on our test bed, indicating the significance of Bayesian updating along with belief-dependent base-stock levels in our problem.

6.2. Performance Evaluation of the Myopic Base-Stock Policy

We now adapt the myopic base-stock policy introduced by Veinott (1965) to our inventory model as a heuristic replenishment policy. In this heuristic, the order quantity in period t is determined according to a myopic belief-dependent base-stock level $\tilde{S}^{\pi^t}$ that is calculated from the newsvendor formula applied to the lead-time demand distribution conditional on the current belief π^t, i.e.,

$$\tilde{S}^{\pi^t} = \underset{k \in \{0, \ldots, M\}}{\arg\min} \left(\mathbb{P}\left\{ \sum_{n=0}^{l} w_{t+n} \leq k \,\middle|\, \pi^t \right\} \geq \frac{b}{h+b} \right).$$

Figure 4 95% confidence intervals for $100 \times \frac{\tilde{\lambda}-\lambda^*_{64}}{\lambda^*_{64}}$ vs. p when $b = 20,\ l \in \{0,1,2\},\ P \in \{P_1, P_2, P_3\},\ h \in \{2,5,10\}$.



For each instance, we simulate the inventory system under this myopic belief-dependent base-stock policy, calculating the average cost denoted by $\tilde{\lambda}$ in each replication. Figure 4 exhibits the 95% confidence intervals for the percentage difference from the optimal average cost λ^*_{64}, i.e., $100 \times \frac{\tilde{\lambda}-\lambda^*_{64}}{\lambda^*_{64}}$.

We observe from Figure 4 that the confidence intervals contain zero in 92 of the 108 instances: The myopic base-stock policy appears to be optimal at a confidence level of 95% in those instances. We also note that the largest optimality gaps (no more than 3.07%) tend to occur when $p = 0.1$ and $h = 10$: The myopic base-stock policy is optimal if $y_t \leq \tilde{S}^{\pi^t}$ with probability one (Lovejoy 1992). It performs worse as the likelihood of excess inventory at the beginning of any period, i.e., $\mathbb{P}\{y_t \geq \tilde{S}^{\pi^t}\}$, increases. For the instances with $p = 0.1$, in a single period, the lowest possible expected demand is $20 \times p = 2$ while the highest possible expected demand is $20 \times (1-p) = 18$. For such instances with highly fluctuating demand, the base-stock levels are likely to vary more widely over time, leading to a larger $\mathbb{P}\{y_t \geq \tilde{S}^{\pi^t}\}$. Hence, and since the holding cost is high, a worse performance results.

In the literature Treharne and Sox (2002) have evaluated the performance of the myopic belief-dependent base-stock policy in the *finite-horizon total-cost* problem (with Markov-modulated demand and partial information). They consider instances with three demand states in which the demand distributions are negative binomial with means 1, 8, and 16, respectively. They rescale their distributions by truncating them at a maximum demand of 18. The transition matrices, holding and shortage costs, and lead times in our instances are similar to those in only some of their instances. On a test bed of 252 instances, for the myopic policy, they have found that the average optimality gap is 5.19% and the largest optimality gap is 44.84%. This poor performance of the myopic policy reported in Treharne and Sox (2002) prompted us to extend our numerical experiments in the *average-cost* problem (again with Markov-modulated demand and partial information) to include the compiled instances in Treharne and Sox (2002). On the same test bed, for the myopic policy, we have found that the average optimality gap is only 0.41% and the largest optimality gap is 3.61%. We thus conclude that the myopic policy performs significantly better in the average-cost problem than in the finite-horizon total-cost problem.

7. Concluding Remarks

We have studied the inventory replenishment problem when the demand distribution undergoes Markovian transitions over time. The state of the underlying Markov chain can only be partially observed based on past demand data. After formulating this problem as an MDP with Bayesian updating, we establish the optimality of a belief-dependent base-stock policy in the discounted-cost case. Using the vanishing discount method, when the underlying Markov chain is ergodic, we extend the optimality of the belief-dependent base-stock policy to the average-cost case. Numerical experiments have revealed the outstanding average-cost performance of the myopic belief-dependent base-stock policy. Structural insights gained from our study further our understanding of data-driven approaches in inventory management that involve Bayesian updating.

Future extensions of this paper could consider inventory models with fixed replenishment order costs. In the literature dealing with fixed ordering costs, Beyer and Sethi (1997) have shown the optimality of a state-dependent (s, S) policy for average-cost inventory models with perfectly observed Markov-modulated demand. Leveraging our structural analysis, the optimality of (s, S) policies may be extended to average-cost models with partially observed Markov-modulated demand. Our research may also guide future research aimed at characterizing the optimal policy structure in more complex average-cost inventory models, such as multi-item and/or multi-echelon inventory systems with partial demand information. Lastly, future research could study the inventory replenishment problem under more limited information about demand. Examples include inventory models with unknown demand distributions and unknown transition matrices for the underlying Markov chain, and inventory models with unknown numbers of demand states. The Baum-Welch and Viterbi algorithms may be employed in estimation of such unknown parameters, enabling good approximations of the original problem. See Mamani et al. (2017), Ban and Rudin (2019), Xin and Goldberg (2019), and Zhang et al. (2019) for recent work on data-driven approaches in the inventory literature.

Acknowledgments

The authors thank Jayashankar Swaminathan, the senior editor, and two anonymous referees for their constructive comments and suggestions. The authors are also grateful to Serdar Yüksel at Queen's University for his helpful comments on an earlier version of the paper. This research was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) [Grant 214M243].

References

Arifoğlu K, Özekici S (2010) Optimal policies for inventory systems with finite capacity and partially observed Markov-modulated demand and supply processes. *European Journal of Operational Research* 204(3):421–438.

Arifoğlu K, Özekici S (2011) Inventory management with random supply and imperfect information: A hidden Markov model. *International Journal of Production Economics* 134(1):123–137.

Azoury KS (1985) Bayes solution to dynamic inventory models under unknown demand distribution. *Management Science* 31(9):1150–1160.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Barber D (2012) *Bayesian reasoning and machine learning* (Cambridge University Press).

Bayraktar E, Ludkovski M (2010) Inventory management with partially observed nonstationary demand. *Annals of Operations Research* 176(1):7–39.

Bertsekas DP (2012) *Dynamic Programming and Optimal Control, Vols. II* (Athena Scientific).

Bertsekas DP (2017) *Dynamic Programming and Optimal Control, Vols. I* (Athena Scientific).

Beyer D, Cheng F, Sethi SP, Taksar M (2010) *Markovian demand inventory models* (Springer).

Beyer D, Sethi SP (1997) Average cost optimality in inventory models with Markovian demands. *Journal of Optimization Theory and Applications* 92(3):497–526.

Borkar VS (2000) Average cost dynamic programming equations for controlled Markov chains with partial observations. *SIAM Journal on Control and Optimization* 39(3):673–681.

Chao X, Gong X, Shi C, Yang C, Zhang H, Zhou SX (2018) Approximation algorithms for capacitated perishable inventory systems with positive lead times. *Management Science* 64(11):5038–5061.

Chen L (2010) Bounds and heuristics for optimal bayesian inventory control with unobserved lost sales. *Operations Research* 58(2):396–413.

Chen L, Mersereau AJ (2015) Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. *Retail Supply Chain Management*, 79–112 (Springer).

Ding X, Puterman ML, Bisi A (2002) The censored newsvendor and the optimal acquisition of information. *Operations Research* 50(3):517–527.

Feinberg EA, Kasyanov PO, Zadoianchuk NV (2012) Average cost Markov decision processes with weakly continuous transition probabilities. *Mathematics of Operations Research* 37(4):591–607.

Fernández-Gaucherand E, Arapostathis A, Marcus SI (1991) On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision process. *Annals of Operations Research* 29(1):439–470.

Graves SC, Meal HC, Dasu S, Qui Y (1986) Two-stage production planning in a dynamic environment. *Multi-stage Production Planning and Inventory Control*, 9–43 (Springer).

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts ITH application to safety stock analysis in production/distribution systems. *IIE Transactions* 26(3):17–30.

Heese HS, Swaminathan JM (2010) Inventory and sales effort management under unobservable lost sales. *European Journal of Operational Research* 207(3):1263–1268.

Hu J, Zhang C, Zhu C (2016) (s, S) inventory systems with correlated demands. *INFORMS Journal on Computing* 28(4):603–611.

Huh WT, Janakiraman G, Nagarajan M (2011) Average cost single-stage inventory models: An analysis using a vanishing discount approach. *Operations Research* 59(1):143–155.

Iglehart DL (1964) The dynamic inventory problem with unknown demand distribution. *Management Science* 10(3):429–440.

Johnson GD, Thompson H (1975) Optimality of myopic inventory policies for certain dependent demand processes. *Management Science* 21(11):1303–1307.

Kaminsky P, Swaminathan JM (2001) Utilizing forecast band refinement for capacitated production planning. *Manufacturing & Service Operations Management* 3(1):68–81.

Kaminsky P, Swaminathan JM (2004) Effective heuristics for capacitated production planning with multiperiod production and demand with forecast band refinement. *Manufacturing & Service Operations Management* 6(2):184–194.

Karlin S (1960) Dynamic inventory policy with varying stochastic demands. *Management Science* 6(3):231–258.

Kesavan S, Kushwaha T (2014) Differences in retail inventory investment behavior during macroeconomic shocks: Role of service level. *Production and Operations Management* 23(12):2118–2136.

Lariviere MA, Porteus EL (1999) Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science* 45(3):346–363.

Lovejoy WS (1990) Myopic policies for some inventory models with uncertain demand distributions. *Management Science* 36(6):724–738.

Lovejoy WS (1991) Computationally feasible bounds for partially observed Markov decision processes. *Operations Research* 39(1):162–175.

Lovejoy WS (1992) Stopped myopic policies in some inventory models with generalized demand processes. *Management Science* 38(5):688–707.

Malladi SS, Erera AL, White III CC (2019) Inventory control with modulated demand and a partially observed modulation process. Working paper.

Mamani H, Nassiri S, Wagner MR (2017) Closed-form solutions for robust inventory management. *Management Science* 63(5):1625–1643.

Miller BL (1986) Scarf's state reduction method, flexibility, and a dependent demand inventory model. *Operations Research* 34(1):83–90.

Natarajan KV, Swaminathan JM (2014) Inventory management in humanitarian operations: Impact of amount, schedule, and uncertainty in funding. *Manufacturing & Service Operations Management* 16(4):595–603.

Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. *Management Science* 57(6):1111–1137.

Puterman ML (2014) *Markov decision processes: Discrete stochastic dynamic programming* (John Wiley & Sons).

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rhenius D (1974) Incomplete information in Markovian decision models. *The Annals of Statistics* 1327–1334.

Ross SM (1968) Arbitrary state Markovian decision processes. *The Annals of Mathematical Statistics* 39(6):2118–2122.

Saldı N, Yüksel S, Linder T (2017) On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Mathematics of Operations Research* 42(4):945–978.

Sandıkçı B (2010) Reduction of a POMDP to an MDP. *Wiley Encyclopedia of Operations Research and Management Science* .

Scarf H (1959) Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics* 30(2):490–508.

Scarf HE (1960) Some remarks on Bayes solutions to the inventory problem. *Naval Research Logistics Quarterly* 7(4):591–596.

Sethi SP, Cheng F (1997) Optimality of (s,S) policies in inventory models with Markovian demand. *Operations Research* 45(6):931–939.

Shamir N, Shin H (2016) Public forecast information sharing in a market with competing supply chains. *Management Science* 62(10):2994–3022.

Shang KH (2012) Single-stage approximations for optimal policies in serial inventory systems with nonstationary demand. *Manufacturing & Service Operations Management* 14(3):414–422.

Song J, Zipkin P (1993) Inventory control in a fluctuating demand environment. *Operations Research* 41(2):351–370.

Spiliotopoulou E, Donohue K, Gürbüz MÇ (2016) Information reliability in supply chains: the case of multiple retailers. *Production and Operations Management* 25(3):548–567.

Treharne JT, Sox CR (2002) Adaptive inventory control for nonstationary demand and partial information. *Management Science* 48(5):607–624.

Veinott AF (1965) Optimal policy for a multi-product, dynamic, nonstationary inventory problem. *Management Science* 12(3):206–222.

Xin L, Goldberg DA (2019) Distributionally robust inventory control when demand is a martingale. Working paper.

Yu H, Bertsekas DP (2004) Discretized approximations for POMDP with average cost. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 619–627.

Zhang H, Chao X, Shi C (2019) Closing the gap: A learning algorithm for the lost-sales inventory system with lead times. *Management Science* Forthcoming.

Zhou R, Hansen EA (2001) An improved grid-based approximation algorithm for POMDPs. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, 707–716.