Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Learning in Structured MDPs with Convex Cost Functions: Improved Regret Bounds for Inventory Management

Shipra Agrawal, Randy Jia

To cite this article:
Shipra Agrawal, Randy Jia (2022) Learning in Structured MDPs with Convex Cost Functions: Improved Regret Bounds for Inventory Management. Operations Research 70(3):1646-1664. https://doi.org/10.1287/opre.2022.2263

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 3, May–June 2022, pp. 1646–1664
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Learning in Structured MDPs with Convex Cost Functions: Improved Regret Bounds for Inventory Management

Shipra Agrawal,[a] Randy Jia[a,*]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
*Corresponding author

Contact: sa3305@columbia.edu, https://orcid.org/0000-0003-4486-3871 (SA); rqj2000@columbia.edu, https://orcid.org/0000-0002-7101-9572 (RJ)

Received: July 31, 2019
Revised: February 22, 2021
Accepted: December 21, 2021
Published Online in Articles in Advance: March 25, 2022

Area of Review: Market Analytics and Revenue Management

https://doi.org/10.1287/opre.2022.2263

Abstract. We consider a stochastic inventory control problem under censored demand, lost sales, and positive lead times. This is a fundamental problem in inventory management, with significant literature establishing near optimality of a simple class of policies called "base-stock policies" as well as the convexity of long-run average cost under those policies. We consider a relatively less studied problem of designing a learning algorithm for this problem when the underlying demand distribution is unknown. The goal is to bound the regret of the algorithm when compared with the best base-stock policy. Our main contribution is a learning algorithm with a regret bound of $\tilde{O}((L+1)\sqrt{T}+D)$ for the inventory control problem. Here, $L \geq 0$ is the fixed and known lead time, and D is an unknown parameter of the demand distribution described roughly as the expected number of time steps needed to generate enough demand to deplete one unit of inventory. Notably, our regret bounds depend linearly on L, which significantly improves the previously best-known regret bounds for this problem where the dependence on L was exponential. Our techniques utilize the convexity of the long-run average cost and a newly derived bound on the "bias" of base-stock policies to establish an almost black box connection between the problem of learning in Markov decision processes (MDPs) with these properties and the stochastic convex bandit problem. The techniques presented here may be of independent interest for other settings that involve large structured MDPs but with convex asymptotic average cost functions.

Funding: This research was supported in part by an Amazon Research Award 2017, awarded to author Shipra Agrawal.

Keywords: inventory control problem • censored demand • reinforcement learning • online convex optimization • regret bounds

1. Introduction

Many operations management problems involve making decisions sequentially over time, where the outcome of a decision may depend on the current state of the system in addition to an uncertain demand or customer arrival process. This includes several online decision-making problems in revenue and supply chain management. There, the sales revenue and supply costs incurred as a result of pricing and ordering decisions may depend on the current level of inventory in stock, back orders, outstanding orders, etc. in addition to the uncertain demand and/or supply for the products. A Markov decision process (MDP) is a useful framework for modeling these sequential decision-making problems. In a typical formulation, the state of the MDP captures the current position of inventory. The reward (observed sales) depends on the current state of the inventory in addition to the demand. The stochastic state transition and reward generation models capture the uncertainty in demand.

A fundamental yet notoriously difficult problem in this area is the periodic inventory control problem under *positive lead times* and *lost sales* (Zipkin 2000, 2008). In this problem, in each of the T sequential decision-making periods, the decision maker takes into account the current on-hand inventory and the pipeline of outstanding orders to decide the new order. There is a fixed delay (i.e., lead time) between placing an order and receiving it. A random demand is generated from a static distribution, independently in every period. However, the demand information is *censored* in the sense that the decision maker observes only the sales (i.e., minimum of the demand and the on-hand inventory). Any unmet demand is lost and incurs a penalty called the lost sales penalty. Any left-over inventory at the end of a period incurs a holding cost. The aim is to minimize the aggregate long-term inventory holding cost and lost sales penalty. There is a significant existing research that develops a Markov model (or semi-Markov model as the lost sales penalty is unobserved) for this problem and studies methods for computing optimal policies, assuming the demand distribution is *either known or can be efficiently*

simulated (e.g., see the survey in Bijvank and Vis 2011). In particular, a simple class of policies called base-stock policies has been shown to be theoretically near optimal for this problem (Huh et al. 2009b, Bijvank and Vis 2011): that is, the cost of the optimal base-stock policy converges to the cost of the optimal policy as the lost sales penalty grows. Under a base-stock policy, the inventory position is always maintained at a target "base-stock" level. Notably, when using a base-stock policy, the infinite horizon average cost function for the inventory control MDP can be shown to be convex in the base-stock level (Janakiraman and Roundy 2004). Therefore, under the known demand model, convex optimization can be used to compute the optimal base-stock policy.

In this paper, we considered a relatively less studied problem of periodic inventory control when the decision maker does not know the demand distribution a priori. The goal is to design a learning algorithm that can use the observed outcomes of past decisions to implicitly learn the unknown underlying MDP model and adaptively improve the decision-making strategy over time (also known as a reinforcement learning algorithm). Because computing an optimal policy for the inventory control problem in general is fundamentally difficult, we will benchmark our algorithm against a more well-understood target: the optimal base-stock policy. Although this is not the true optimal policy, the simplistic nature along with theoretical guarantees of near optimality (see theorem 5 of Huh et al. 2009b) of base-stock policies makes it an attractive benchmark to measure our learning algorithm against.

The two main challenges in designing an efficient learning algorithm for the inventory control problem described are presented by the *censored demand* and the *positive lead time*. The censored demand assumption results in an exploration-exploitation trade-off for the learning algorithm. Because the decision maker can only observe the sales, which is the minimum of the demand and the on-hand inventory for a product, the quality of samples available for demand estimation of a product depends crucially on the past ordering decisions. For example, suppose that because of the past ordering policies, a certain product was maintained at a low inventory level for most of the past sales periods. Then, the higher quantiles of the demand distribution for that product would be unobserved. Therefore, in order to ensure accurate demand learning, large inventory states need to be sufficiently explored. However, this exploration needs to be limited because of the holding cost incurred for any leftover inventory. There has been significant recent work on exploration-exploitation algorithms for regret minimization in finite state and finite action MDPs, with regret bounds that depend linearly or sublinearly on the size of the state space and the action space (e.g., Bartlett and Tewari 2009, Jaksch et al. 2010, Agrawal and Jia 2017). However, the positive lead time in delivery of an order results in a much enlarged state space (exponential in lead time) for the inventory control problem considered here because the state needs to track all the outstanding orders in the pipeline. There is a further issue of discretization because the state space (inventory position) and the action space (orders) are continuous. Discretizing over a grid would result in a further enlarged state space and action space. As a result, none of these reinforcement learning techniques can be applied directly to obtain useful regret bounds for the inventory control problem considered here.

The main insight in this paper is that even though the state space is large, the convexity of the average cost function under the benchmark policies (here, base-stock policies) can be used to design an efficient learning algorithm for this MDP. We use the relation between *bias* and *infinite horizon average cost* of a policy given by Bellman equations to provide a connection between stochastic convex bandit optimization and the problem of learning and optimization in such MDPs. Specifically, we build upon the algorithm for stochastic convex optimization with bandit feedback from Agarwal et al. (2011) to derive a simple algorithm that achieves an $\tilde{O}((L+1)\sqrt{T}+D)$ regret bound for the inventory control problem. Here, L is the fixed and known lead time, and D is a parameter of the demand distribution F, defined as the expected number of independent draws needed from distribution F for the sum to exceed one. Importantly, although our regret bound depends on D, our algorithm does not need to know this parameter. The $\tilde{O}(\cdot)$ notation hides logarithmic factors and absolute constants.

Our regret bound substantially improves the existing results for this problem, provided by Huh et al. (2009a) and Zhang et al. (2020), where the regret bounds grow exponentially with the lead time L (roughly as $D^L\sqrt{T}$), and many further assumptions on the demand distribution are required for the bounds to hold. A more detailed comparison with the related work is provided later in the text. More importantly, we believe that our algorithm design and analysis techniques can be applied in an almost black box manner for minimizing regret in other problem settings involving MDPs whose cost functions are convex under benchmark policies. Such convexity results are available for many other operations management problems: for example, for several formulations of admission control and server allocation problems in queuing (Weber 1980, Lee and Cohen 1983, Shanthikumar and Yao 1987). Therefore, the techniques presented here may be of independent interest.

1.1. Organization

The rest of the paper is organized as follows. In the following subsections, we provide a formal problem

definition and describe our main results along with a precise comparison of our regret bounds with closely related work. In Section 2, we present a learning algorithm and regret analysis for optimizing a general MDP with a convex cost function. We adapt this algorithm design and regret analysis to the inventory management problem in Section 3. We conclude in Section 4.

1.2. Problem Formulation

We consider a single-product stochastic inventory control problem with lost sales and positive lead times. The problem setting considered here is similar to the setting considered in Huh et al. (2009b) and Zhang et al. (2020). An inventory manager makes sequential decisions in discrete time steps $t = 1, \ldots, T$. In the beginning of every time step t, the inventory manager observes the current inventory level inv_t and L previous unfulfilled orders in the pipeline, denoted as $o_{t-L}, o_{t-L+1}, \ldots, o_{t-1}$, for a single product. Here, $L \geq 0$ is the lead time defined as the delay (number of time steps) between placing an order and receiving it. Initially, in step 1, there is no inventory ($\text{inv}_1 = 0$) and no unfulfilled orders. Based on this information, the manager decides the amount $o_t \in \mathbb{R}$ of the product to order in the current time step.

The next inventory position is then obtained through the following sequence of events. First, the order o_{t-L} that was made L time steps earlier arrives, so that the on-hand inventory level becomes $I_t = \text{inv}_t + o_{t-L}$. Then, an *unobserved* demand $d_t \geq 0$ is generated from an unknown demand distribution F, independent of the previous time steps. Sales is the minimum of the on-hand inventory and the demand (i.e., sales $y_t := \min\{I_t, d_t\}$). The decision maker only observes the sales y_t and not the actual demand d_t—the demand information is, therefore, *censored*. A holding cost of $h(I_t - d_t)^+$ is incurred on remaining inventory, and a lost sales penalty of $p(d_t - I_t)^+$ is incurred on the part of demand that could not be served because of insufficient on-hand inventory. That is, the cost incurred at the end of step t is

$$\bar{C}_t = h(I_t - d_t)^+ + p(d_t - I_t)^+, \tag{1}$$

where $(I_t - d_t)^+ = \max(I_t - d_t, 0), (d_t - I_t)^+ = \max(d_t - I_t, 0)$, and h, p are prespecified constants denoting per unit holding cost and per unit lost sales penalty, respectively. Note that the lost sales and therefore, the lost sales penalty are unobserved by the decision maker.

Figure 1 illustrates the described sequence of arrivals of orders and demand. The next step $t + 1$ begins with the leftover inventory

$$\text{inv}_{t+1} : (I_t - d_t)^+ = (\text{inv}_t + o_{t-L} - d_t)^+ \tag{2}$$

and the new pipeline of outstanding orders $o_{t-L+1}, \ldots, o_t$.

An online learning algorithm for this problem needs to sequentially decide the orders $o_1, \ldots, o_T$ under demand censoring and without a priori knowledge of the demand distribution. The objective is to minimize the total expected cost $\mathbb{E}[\sum_{t=1}^{T} \bar{C}_t]$.

Base-stock policies (also known as "order up" to policies) form an important class of policies for the inventory control problem. Under such a policy, the inventory manager always orders a quantity that brings the total inventory position (i.e., the sum of leftover inventory plus outstanding orders) to some fixed value known as the base-stock level, if possible. Specifically, in the beginning of a step t, let the leftover inventory be inv_t and the outstanding orders be $o_{t-L}, \ldots, o_{t-1}$. Then, on using a base-stock policy with level x, the new order o_t in step t is given by $o_t = (x - \text{inv}_t - \sum_{i=1}^{L} o_{t-i})^+$. Zipkin (2008) and Huh et al. (2009b) provide empirical results that show that base-stock policies work well in many applications. Furthermore, Huh et al. (2009b) show that as the ratio of per unit lost sales penalty to holding cost increases to infinity, the ratio of the total cost incurred by the best base-stock policy to the optimal cost converges to one. Because the ratio of per unit lost sales penalty to holding cost is typically large in many applications, the best base-stock policy can be considered close to optimal.

Considering the asymptotic optimality of base-stock policies, several past works consider a more tractable objective of minimizing the regret of an online algorithm compared with the best base-stock policy (e.g., Huh et al. 2009a, Zhang et al. 2020).

1.2.1. Convexity Property.

Let $\bar{C}_t^x, t = 1, 2\ldots,$ denote the sequence of costs incurred on running the base-stock policy with level x. Define λ^x as the expected infinite horizon average cost of this base-stock policy when starting from no inventory and no outstanding orders: that is,

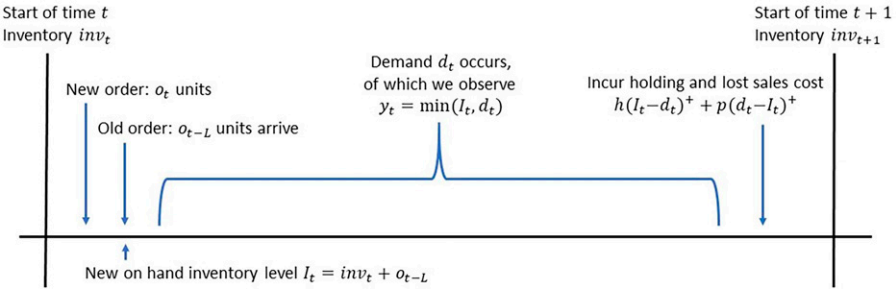
$$\lambda^x := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \bar{C}_t^x \middle| \text{inv}_1 = 0\right]. \tag{3}$$

We also refer to the λ^x as the long-run average cost of this policy. The following result from Janakiraman and Roundy (2004) shows that this long-run average cost is convex in x.

Lemma 1 (Derived from theorem 12 of Janakiraman and Roundy 2004). *Given a demand distribution F, $F(0) > 0$ (i.e., there is a nonzero probability of zero demand). Then, for any $x \geq 0$, the expected infinite horizon average cost, λ^x, of the corresponding base-stock policy is convex in x.*

Remark 1. Theorem 12 of Janakiraman and Roundy (2004) actually proves convexity of expected average cost when starting from inventory level $\text{inv}_1 = x$. However, in the definition of λ^x, we assumed there is no starting inventory (i.e., $\text{inv}_1 = 0$). On starting from no inventory and no outstanding orders and using the

Figure 1. (Color online) Timing of Arrival of Orders and Demand at Time t



base-stock policy with level x, the system will reach the state with inventory level x and no outstanding orders in finite (exactly L) steps. Therefore, λ^x is same as the expected infinite horizon average cost incurred on starting with inventory level $\text{inv}_1 = x$.

1.2.2. Regret Against Any Base-Stock Policy. Regret of an algorithm against any given base-stock policy with level x is defined as

$$\text{Regret}(T, x) := \mathbb{E}\left[\sum_{t=1}^{T} \bar{C}_t\right] - \mathbb{E}\left[\sum_{t=1}^{T} \bar{C}_t^x\right], \tag{4}$$

where $\bar{C}_t, \bar{C}_t^x, t = 1, 2\ldots,$ are the sequence of costs incurred on running the algorithm and the policy with base-stock level x, respectively, starting from no inventory and no outstanding orders. The expectations are taken with respect to any randomness in the algorithm as well as randomness in the demand. Then, given $[0, U]$, a prespecified range of base-stock levels to be considered, we bound worst-case regret:

$$\text{Regret}(T) := \max_{x \in [0, U]} \text{Regret}(T, x). \tag{5}$$

1.3. Main Results

Before we formally state our main result, we define D, a parameter of the demand distribution F that appears in our regret bounds. It is important to note that our algorithm does not need to know the parameter D.

Definition 1. Define D as the expected number of independent samples needed from distribution F for the sum of those samples to exceed one. More precisely, let $d_1, d_2, d_3, \ldots,$ denote a sequence of independent samples generated from the demand distribution F, and let τ be the minimum number such that $\sum_{i=1}^{\tau} d_i \geq 1$. Then, define $D := \mathbb{E}[\tau]$. We refer to D as the expected time to deplete one unit of inventory. We assume that the demand distribution F is such that D is finite.

Our main result is stated as follows.

Theorem 1. *Assuming that demand distribution F is such that $F(0) > 0$ and the expected time D to deplete one unit of inventory is finite, then given any lead time $L \geq 0$, there exists an algorithm (Algorithm 1) for the inventory control problem with regret bounded as*

$$\text{Regret}(T) \leq O\Big(D\max(h,p)U^2\log^2(T) + (L+1)\max(h,p)U\sqrt{T\log^3(T)}\Big).$$

For $T \geq (DU)^2$, this implies a regret bound of

$$\text{Regret}(T) \leq \tilde{O}\Big((L+1)\max(h,p)U\sqrt{T}\Big),$$

where $\tilde{O}(\cdot)$ hides logarithmic factors in h, p, U, L, T, and absolute constants.

Here, constants $\max(h,p)$ and U define the scale of the problem. Note that the regret bound has a very mild (additive) dependence on the parameter D of the demand distribution. We conjecture that such a dependence on D in the regret may be unavoidable because every time a learning algorithm reaches an inventory level higher than the optimal base-stock level, it must necessarily wait for time steps roughly proportional to D for the inventory to deplete in order to play a better policy. Only an algorithm that never overshoots the optimal inventory level may avoid incurring this waiting time. However, without a priori knowledge of the optimal level, an exploration-based learning algorithm is unlikely to avoid this completely. The appearance of D here also reminds of the appearance of diameter D in regret bounds for general finite MDPs, where diameter is defined as the expected time to go from one state to another (e.g., see Tewari and Bartlett 2008, Jaksch et al. 2010, Agrawal and Jia 2017).

Remark 2. The assumption $F(0) > 0$ in the theorem is required only for using the result on convexity of infinite horizon average cost given by theorem 12 of Janakiraman and Roundy (2004) (see Lemma 1). The convexity result can in fact be shown to hold under some alternate conditions, like finite support of demand,

or under sufficient discretization of demand, which would also be sufficient for our results.

1.4. Comparison with Related Work

Some earlier works on exploration-exploitation algorithms for the inventory control problem (Huh and Rusmevichientong 2009, Besbes and Muharremoglu 2013) provide $\tilde{O}(\sqrt{T})$ regret bounds but under zero lead time (Huh and Rusmevichientong 2009) and/or perishable inventory (Besbes and Muharremoglu 2013) assumptions. The inventory control problem considered here is exactly the same as that considered in the recent work by Zhang et al. (2020) and the earlier work by Huh et al. (2009a). Therefore, we provide a precise comparison with the results obtained in those works. Both works execute a base-stock policy over constructed time periods called cycles and employ a gradient estimation to revise the policy for the next cycle. In particular, the simulated cycle-update policy of Zhang et al. (2020) updates policies based on a simulated inventory system running in parallel.

Our result matches the $O(\sqrt{T})$ dependence on T in Zhang et al. (2020), improving on the $O(T^{\frac{2}{3}})$ dependence originally given in Huh et al. (2009a). Further, it can be shown (see Zhang et al. 2020, proposition 1) that for $T > 5$, the expected regret for any learning algorithm in this setting is lower bounded by $\Omega(\sqrt{T})$, and thus, our bound is optimal in T (within logarithmic factors). More importantly, our regret bound scales linearly in L as opposed to the exponential dependence on L in Zhang et al. (2020), which can be traced to more delicate analysis of the cost function and inventory model dynamics (Lemma 5). In particular, we note that replacing Lemma 5 with the analogous bound from the analysis in Zhang et al. (2020) would lead to an exponential dependence on lead time as well. Specifically, the regret bound achieved by Zhang et al. (2020) is of order $\tilde{O}(\max(h,p)^2U^2(1/c)^L\sqrt{T})$. Besides having an exponential dependence on L, it depends on a constant c given by the product of some positive probabilities for demand to take values in certain ranges, which requires several further assumptions on the distribution F (see assumption 1 of Zhang et al. 2020). In comparison, our distribution parameter D is milder and more interpretable, and most notably, it appears as an additive (rather than multiplicative) term in the regret bound.

Among other related work, Bartók et al. (2014), Besbes et al. (2015), and Lugosi et al. (2017) provide $\tilde{O}(\sqrt{T})$ regret bounds for variations of the inventory control problem under *adversarial demand*. However, these works make significant simplifying assumptions such as zero lead time and perishing inventory. Under such assumptions, there is no state dependence across periods, and the problem becomes closer to an online learning problem rather than a reinforcement learning problem. Finally, as discussed earlier, the existing work on finite time regret bounds for reinforcement learning algorithms for general finite state MDPs, such as Bartlett and Tewari (2009), Jaksch et al. (2010), and Agrawal and Jia (2017), would imply a regret bound exponential in the lead time because of the exponential size of the state space.

2. Learning to Optimize an MDP with Convex Cost Function

In this section, we present a learning algorithm and regret analysis for any MDP with a convex cost function. Specifically, we consider the problem of regret minimization in an MDP given a single parametric set of policies. The main structural property assumed regarding the MDP is that the average asymptotic average cost, also known as *loss*, under any given policy in this set is convex in the policy parameter. Given this convexity property and a bound on the bias of the MDP, we present a stochastic convex bandit-based learning algorithm with sublinear regret bounds. In subsequent sections, we demonstrate that the required convexity and bounded bias properties are indeed satisfied by the inventory control MDP in order to derive an efficient algorithm and improved regret bounds of the inventory management problem. However, it is important to note that the results presented in this section are more generally applicable to any MDP (discrete or continuous state space) as long as it satisfies the prescribed convexity and bounded bias properties. We present the results in this section in a self-contained manner; the results presented here can be understood and used without going through the rest of the sections in the paper.

Formally, we are given an MDP $\mathcal{M}$ with state space $\mathcal{S}$ and action space $\mathcal{A}$ as well as a parametric set of policies:[1]

$$\Pi = \{\pi^x : \mathcal{S} \to \mathcal{A}, x \in \mathcal{X}\}. \tag{6}$$

A learning algorithm needs to make sequential decisions using one of the policies in Π at every discrete time step $t = 1, \ldots, T$. At every time step t, the algorithm observes current state $s_t \in \mathcal{S}$, chooses $x_t \in \mathcal{X}$, takes action $\pi^{x_t}(s_t)$, and then, observes the cost C_t.

We aim to minimize regret of the algorithm against the policies in Π. Given a starting state $s_1 \in \mathcal{S}$, regret in time T against any policy $\pi^x \in \Pi$ is defined as

$$\text{Regret}(T,x) := \mathbb{E}\left[\sum_{t=1}^{T} C_t \middle| \mathbf{s}_1\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^x \middle| \mathbf{s}_1\right], \tag{7}$$

with C_t^x being the cost incurred at time t on using the policy π^x at all time steps starting from state $\mathbf{s}_1$. Then,

$$\text{Regret}(T) := \max_{x \in \mathcal{X}} \text{Regret}(T, x).$$

We also consider "pseudoregret" against the asymptotic average cost

$$g^x(\mathbf{s}_1) := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} C_t^x | \mathbf{s}_1\right]$$

of policy π^x:

$$\text{Pseudoregret}(T, x) := \mathbb{E}\left[\sum_{t=1}^{T} C_t \middle| \mathbf{s}_1\right] - T g^x(\mathbf{s}_1). \tag{8}$$

We will focus on bounding the pseudoregret for our algorithm against any policy in Π and show that it can be used to derive a similar bound on regret.

We first present some key definitions and results that will be used throughout the algorithm design and regret analysis. In the MDP literature, the stochastic process obtained on fixing a policy in an MDP is referred as a Markov reward process (MRP) (Puterman 2014), which is essentially a Markov chain with a reward (or cost) associated with each state. We denote by $\mathcal{M}(x, s_1)$ the MRP obtained on fixing the policy as π^x for $x \in \mathcal{X}$.

Definition 2 (Markov Reward Process $\mathcal{M}(\boldsymbol{x}, \mathbf{s}_1)$). Given any x and $\mathbf{s}_1 \in \mathcal{S}$, the MRP $\mathcal{M}(x, \mathbf{s}_1)$ is defined by the bipartite stochastic process

$$\{(\mathbf{s}_t, C^x(\mathbf{s}_t)); t = 1, 2, 3, \dots\},$$

where $\mathbf{s}_t$ is distributed according to $P^x(\mathbf{s}_t) := \Pr(s_t | a_t = \pi^x(s_{t-1}))$, the state transition distribution under policy π^x, and $C^x(s_t)$ is the expected cost on taking action $a_t = \pi^x(s_t)$ in state s_t in MDP $\mathcal{M}$. Further, we define $\mathcal{S}^x \subseteq S$ as the set of reachable states in this MRP, when starting from state $\mathbf{s}_1$, so that $\mathbf{s}_t \in \mathcal{S}^x$ for all t.

Two important quantities are the *loss* and *bias* of this MRP.

Definition 3 (Loss and Bias). For any $\mathbf{s} \in \mathcal{S}^x$, the loss $g^x(\mathbf{s})$ of MRP $\mathcal{M}(x, \mathbf{s})$ is the long-run average cost (starting from state $\mathbf{s}$), and the bias $v^x(\mathbf{s})$ of $\mathcal{M}(x, \mathbf{s})$ is the total difference in the cost from the asymptotic average cost (starting from state $\mathbf{s}$). More formally, define (assuming limits exist)

$$g^x(\mathbf{s}) := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} C^x(\mathbf{s}_t) | \mathbf{s}_1 = \mathbf{s}\right]$$

$$v^x(\mathbf{s}) := \mathbb{E}\left[\lim_{T\to\infty} \sum_{t=1}^{T} C^x(\mathbf{s}_t) - g^x(\mathbf{s}_t) | \mathbf{s}_1 = \mathbf{s}\right].$$

The following relation between loss and bias is known.

Lemma 2 (Puterman 2014, theorem 8.2.6). *For any state $\mathbf{s} \in \mathcal{S}^x$ in MRP $\mathcal{M}(x, \mathbf{s}_1)$, the bias and loss satisfy the following equation*:

$$g^x(\mathbf{s}) = C^x(\mathbf{s}) + \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s})}[v^x(\mathbf{s}')] - v^x(\mathbf{s}),$$

where $P^x(\mathbf{s})$ was defined as the probability distribution of next state given state $\mathbf{s}$.

A main new technical result that we derive and utilize in this work is the following concentration lemma for any MRP with bounded bias.

Lemma 3 (Concentration Under Bounded Bias). *Assume we are given an MRP $\mathcal{M}(x, \mathbf{s}_1)$ such that the gain is uniform across states, i.e.,*

$$g^x(\mathbf{s}) = g(\mathbf{s}') =: g^x, \ \forall \mathbf{s}, \mathbf{s}',$$

and the span of the bias v^x is bounded by H: that is,

$$|v^x(\mathbf{s}) - v^x(\mathbf{s}')| \le H, \ \forall \mathbf{s}, \mathbf{s}' \in \mathbf{s}^x.$$

Let $\mathbf{s}_1, \dots, \mathbf{s}_N$ denote an observed sequence of first N states generated by MRP $\mathcal{M}(x, \mathbf{s}_1)$. Then, for any $\delta > 0$, with probability $1 - \delta$,

$$\left|\frac{1}{N}\sum_{t=1}^{N} C^x(\mathbf{s}_t) - g^x(\mathbf{s}_1)\right| \le \frac{H}{N} + H\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Proof. By Lemma 2, the loss g^x and bias v^x for policy π^x satisfy $g^x(\mathbf{s}) = C^x(\mathbf{s}) + \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s})}[v^x(\mathbf{s}')] - v^x(\mathbf{s})$ for all states $\mathbf{s} \in \mathcal{S}^x$. Also, $g^x = g^x(\mathbf{s}_1) = g^x(\mathbf{s}_t)$ for all t. We use these observations to derive the following:

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{t=1}^{N} C^x(\mathbf{s}_t)\right) - g^x(\mathbf{s}_1)\right| = \left|\frac{1}{N}\sum_{t=1}^{N}\left(C^x(\mathbf{s}_t) - g^x(\mathbf{s}_t)\right)\right| \\
&= \left|\frac{1}{N}\sum_{t=1}^{N}\left(C^x(\mathbf{s}_t) - \left(C^x(\mathbf{s}_t) + \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_t)}[v^x(\mathbf{s}')] - v^x(\mathbf{s}_t)\right)\right)\right| \\
&= \left|\frac{1}{N}\sum_{t=1}^{N}\left(v^x(\mathbf{s}_t) - \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_t)}[v^x(\mathbf{s}')]\right)\right| \\
&= \left|\frac{1}{N}\left(v^x(\mathbf{s}_1) - \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_N)}[v^x(\mathbf{s}')]\right)\right. \\
&\quad \left. + \frac{1}{N}\sum_{t=1}^{N-1}\left(v^x(\mathbf{s}_{t+1}) - \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_t)}[v^x(\mathbf{s}')]\right)\right| \\
&\le \frac{H}{N} + \left|\frac{1}{N}\sum_{t=1}^{N-1}\left(v^x(\mathbf{s}_{t+1}) - \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_t)}[v^x(\mathbf{s}')]\right)\right|.
\end{aligned}$$

For the last inequality, we used the assumed bound on bias. Now, let

$$\Delta_{t+1} := v^x(\mathbf{s}_{t+1}) - \mathbb{E}_{\mathbf{s}' \sim P^x(\mathbf{s}_t)}[v^x(\mathbf{s}')].$$

Note that $\mathbb{E}[\Delta_{t+1} | s_t] = 0$ and $\Delta_{t+1} \le H$ (because of the assumed bias bound). Therefore, $\Delta_{t+1}, t = 1, 2, \dots$, is a bounded martingale difference sequence. We apply the Azuma–Hoeffding inequality (refer to Lemma A.1 in Appendix A) to obtain that, for any $\epsilon > 0$,

$$P\left(\left|\sum_{t=2}^{N} \Delta_t\right| \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{2(N-1)H^2}\right).$$

Therefore, by setting $\epsilon = H\sqrt{2(N-1)\log(2/\delta)}$, we obtain that with probability at least $1-\delta$, $\left|\sum_{t=2}^{N}\Delta_t\right| \le H\sqrt{2(N-1)\log(2/\delta)}$. Substituting, we obtain

$$\left|\frac{1}{N}\sum_{t=1}^{N} C^x(\mathbf{s}_t) - g^x(\mathbf{s}_1)\right| \le \frac{H}{N} + H\sqrt{\frac{2\log(2/\delta)}{N}}.$$

Similarly, we can show concentration of average observed cost versus asymptotic average cost.

Lemma 4 (Concentration of Average Observed Versus Expected Cost). *Given MRP $\mathcal{M}(x,\mathbf{s}_1)$ with uniform gain and bias bounded by H, as defined in Lemma 3, let C_t^x be observed cost such that $\mathbb{E}[C_t^x|\mathbf{s}_t] = C^x(\mathbf{s}_t)$. Assume that C_t^x is bounded between by $C_{\max}$. Then, for any $\mathbf{s}_1 \in \mathcal{S}$, with probability $1-\delta$,*

$$\left|\frac{1}{N}\sum_{t=1}^{N} C_t^x - g^x(\mathbf{s}_1)\right| \le \frac{H}{N} + (C_{\max}+H)\sqrt{\frac{2\log(4/\delta)}{N}}.$$

Proof. Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t$ be the sequence of states observed in MRP $\mathcal{M}(x,\mathbf{s}_1)$ starting from state $\mathbf{s}_1$, and let $\mathcal{F}_t$ be the filtration with respect to those states $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t$. Define $X_t = C^x(\mathbf{s}_t) - C_t^x$. Note that $|X_t| \le |C^x(\mathbf{s}_t)| + |C_t^x| \le 2C_{\max}$ and $\mathbb{E}[X_t|F_t] = 0$. Therefore, X_t's form a martingale difference sequence, and a simple application of the Azuma–Hoeffding inequality (Lemma A.1) gives that, with probability $1-\delta/2$,

$$\frac{1}{N}\left|\sum_{t=1}^{N} C^x(\mathbf{s}_t) - \sum_{t=1}^{N} C_t^x\right| \le 2C_{max}\sqrt{\frac{2\log(4/\delta)}{N}}.$$

Combining the inequality with Lemma 3, we get the corollary.

2.1. Algorithm Design

Here, we present a regret minimization algorithm for the case when $\mathcal{X} = [0,U]$ for some $U_{\min}, U_{\max} \in \mathbb{R}$ (i.e., Π is a set of single-parameter policies with the parameter range being a contiguous interval). Further, we assume following properties.

Assumption 1. *We are given an MDP $\mathcal{M}$, policy set $\Pi = \{\pi^x : x \in [0,U]\}$, and a starting state $\mathbf{s}_1$ such that for all policies in Π, the MRP $\mathcal{M}(x,\mathbf{s}_1)$ satisfies the following properties*:

a. *the cost function $g^x(\mathbf{s})$ is uniform (i.e., $g^x(\mathbf{s}) = g^x$) for all $\mathbf{s}$,*

b. *the cost function g^x is convex in x and Lipschitz continuous with Lipschitz constant β, and*

c. *the span of bias is bounded by H (i.e., $|v^x(\mathbf{s}) - v^x(\mathbf{s}')| \le H$, $\forall \mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$).*

Utilizing the convexity property, our algorithm design builds upon ideas from exploration-exploitation algorithms for stochastic convex bandits. In particular, because we restrict to the single parametric policies, we extend the algorithm in Agarwal et al. (2011) for one-dimensional stochastic convex bandits.

In the stochastic convex bandit problem, in every round the decision maker chooses a decision x_t and incurs cost $c_t = f(x_t) + y_t$, where f is some fixed but unknown convex function and the noise y_t is zero mean and independent and identically distributed across rounds $t = 1, \ldots, T$. The goal of an online algorithm is to use past observations to make decisions $x_t, t = 1, \ldots, T$ in order to minimize the regret against the best single decision, defined as $\sum_{t=1}^{T}(c_t - f(x^*))$ with $x^* = \arg\min_{x\in\mathbb{X}} f(x)$. Therefore, based on the definition of regret (refer to Equation (7)), one may want to consider a mapping of the problem of regret minimization in MDP to the stochastic convex bandit problem by setting $f(x)$ as $\lambda^x = g^x(\mathbf{s}_1)$, the long-run average cost for policy π^x, which is assumed to be convex in x.

However, a main challenge here is that the instantaneous cost $C^x(\mathbf{s}_t)$ incurred on time t on playing a policy π^{x_t} depends on the current state $\mathbf{s}_t$, and therefore, unlike c_t, $C^x(\mathbf{s}_t)$ is not an unbiased realization of $f(x_t) = \lambda^{x_t}$ (more precisely, the noise is not zero mean and independent and identically distributed across rounds). We overcome this challenge using the concentration result derived in Lemma 3. This concentration result allows us to develop confidence intervals on estimates of cost functions in a manner similar to the stochastic convex bandit algorithms.

Our algorithm, summarized as Algorithm 1, is derived from the algorithm in Agarwal et al. (2011) for one-dimensional stochastic convex bandits. Following are the main components of our algorithm.

2.1.1. Working Intervals of Policy Parameter. The algorithm maintains a confidence interval that contains an optimal value of the policy parameter x, with high probability. Initially, this is set as $[0,U]$, the prespecified interval received as input to the algorithm. As the algorithm progresses, this working interval is refined by discarding portions of the interval, which have low probability of containing an optimal value.

2.1.2. Epoch and Round Structure. The algorithm proceeds in epochs $k = 1, 2, \ldots$. Each epoch is a group of consecutive time steps. A fixed working interval of base-stock levels is maintained throughout an epoch, denoted as $[l_k, r_k]$. Each epoch is further split into groups of consecutive time steps called rounds. The algorithm plays policies $\pi^{x_l}, \pi^{x_c}, \pi^{x_r}$, policies corresponding to parameters $x_l := (l_k + r_k - l_k)/4, x_c := l_k + (r_k - l_k)/2$, $x_r := l_k + (3(r_k - l_k))/4$, respectively. Each of these three policies is played repeatedly for N_i time steps, where $N_i = \log(T)/\gamma_i^2$ with $\gamma_i = 2^{-i}$.

Therefore, the number of sample observations quadruples in each round. At the end of every round, these observations are used to update a confidence interval estimate for the long-run average cost as described next. An epoch ends when the confidence intervals at the end of a round meet a certain condition.

2.1.3. Optional Step: Using π_{init} to Reach a Desired Set of States. We also consider the case when Assumption 1 does not hold for all starting states but only for certain starting states $\mathbf{s}_1 \in \mathcal{S}^x_{\text{init}}$. Additionally, we are given a policy π_{init} to reach a state in $\mathcal{S}^x_{\text{init}}$ in finite expected time D_{init}. In this case, the algorithm uses π^x_{init} to reach one of these desired starting states before playing a policy π^x in every round. If no such policy π^x_{init} is provided, the algorithm skips this step. This relaxed setting will be of use for the inventory control problem, where we will bound bias of a policy π^x only for the MRP obtained on starting from an inventory position where the sum of the on-hand inventory and the outstanding orders is x. Such a state can be easily reached by ordering nothing for some time and then ordering x. In the inventory control problem, D_{init}, the expected time to reach the desired starting state, will be bounded by $L + DU$, with D being the expected time for enough demand to be generated in order to deplete one unit of inventory.

2.1.4. Updating Confidence Intervals. Given a vector $\mathbf{C}_N = (C_1, C_2, \ldots, C_N)$ of observed costs and $\gamma = \sqrt{\log(T)/N}$, define

$$LB(\mathbf{C}_N) := \frac{1}{N}\sum_{i=1}^{N} C_i - 3H\gamma, \text{ and } UB(\mathbf{C}_N) : = \frac{1}{N}\sum_{i=1}^{N} C_i + 3H\gamma, \tag{9}$$

where the bound H on bias is an input to the algorithm.

Now, let $\mathbf{C}_N, \mathbf{C}^c_N, \mathbf{C}^r_N$ denote the $n = N_i$ realizations of pseudocosts (C^x_t) observed on running base-stock policy π^x for each of the three levels $x \in [x_l, x_c, x_r]$ in round i. Then, at the end of round i, the algorithm computes three intervals:

$$[LB(\mathbf{C}^a_N), UB(\mathbf{C}^a_N)] \text{ for } a \in \{l, c, r\}.$$

We show later that for each of these three policies, $g^{x_a} \in [LB(\mathbf{C}^a_N), UB(\mathbf{C}^a_N)]$ with probability $1 - 1/T^2$. This uses Lemmas 3 and 4 proven earlier in order to bound the difference between finite time average and asymptotic average of costs. Therefore, each of these intervals is a high confidence interval for the respective loss.

Algorithm 1 (Learning Algorithm for MDP with Convex Cost Function)

Inputs: Set of policies $\Pi = \{\pi^x : x \in [0, U]\}$, bias bound H, time horizon T.
Optional input: A description of the desirable set of starting states $\mathcal{S}^x_{\text{init}}$ for any policy x and a policy π^x_{init} to reach that set.
Initialize: $[l_1, r_1] := [0, U]$.
for *epochs* $k = 1, 2, \ldots,$ **do**
  Set $w_k := r_k - l_k$, the width of the working interval $[l_k, r_k]$.
  Set $x_l := l_k + w_k/4$, $x_c := l_k + w_k/2$, and $x_r := l_k + 3w_k/4$.
  for *round* $i = 1, 2, \ldots,$ **do**
    Let $\gamma_i = 2^{-i}$ and $N = \log(T)/\gamma_i^2$.
    for $a \in \{l, c, r\}$ **do**
      Play policy $\pi^{x_a}_{\text{init}}$ (if provided) until a time step t such that $\mathbf{s}_t \in \mathcal{S}^{x_a}_{\text{init}}$.
      Play policy π^{x_a} for N time steps to observe N realizations of costs; store as vectors $\mathbf{C}^a_N$.
      If at any point during these two steps, the total number of time steps reaches T, exit.
      Compute confidence interval $[LB(\mathbf{C}^a_N), UB(\mathbf{C}^a_N)]$ of length $6H\gamma_i$, as given by (9).
    end
    if $\max\{LB(\mathbf{C}^l_N), LB(\mathbf{C}^r_N)\} \geq \min\{UB(\mathbf{C}^l_N), UB(\mathbf{C}^c_N), UB(\mathbf{C}^r_N)\} + 6H\gamma_i$ **then**
      if $LB(\mathbf{C}^l_N) \geq LB(\mathbf{C}^r_N)$ **then** $l_{k+1} := x_l$ and $r_{k+1} = r_k$.
      if $LB(\mathbf{C}^l_N) < LB(\mathbf{C}^r_N)$ **then** $l_{k+1} := l_k$ and $r_{k+1} = x_r$
      Go to next epoch $k + 1$.
    else
      Go to next round $i + 1$.
    end if
  end
end

At the end of every round i of an epoch k, the algorithm uses the updated confidence intervals to check if either the portion $[l_k, x_l]$ or the portion $[x_r, r_k]$ of the working interval $[l_k, r_k]$ can be eliminated. Given the confidence intervals, the test used for this purpose is exactly the same as in Agarwal et al. (2011) and uses convexity properties of the loss g^x. If the test succeeds, at least 1/4 of the working interval is eliminated, and the epoch k ends.

2.2. Regret Bound

We prove the following regret bound for Algorithm 1 when applied to any MDP under Assumption 1.

Theorem 2. *Given any MDP $\mathcal{M}$ and a single parametric policy set $\Pi = \{\pi^x : x \in [0, U]\}$, Assumption 1 is satisfied for all starting states $\mathbf{s}_1 \in \mathcal{S}$. Then, the regret of Algorithm 1 is bounded as*

$$Regret(T) = O\left((H + \beta U)\sqrt{T\log^3(\beta UT/H)}\right) = \tilde{O}((\beta U + H)\sqrt{T}).$$

Here, β is the Lipschitz factor for function g^x, and H is the given bound on bias, as per Assumption 1.

We also consider the case when Assumption 1 does not hold for all starting states but only for starting states $\mathbf{s}_1 \in \mathcal{S}_{\text{init}}$. Also, we are given a policy π_{init} to reach a state in $\mathcal{S}_{\text{init}}$ in expected time D_{init}. In this case, Algorithm 1 achieves the following regret bound.

Theorem 3. *Given any MDP $\mathcal{M}$ and a single parametric policy set $\Pi = \{\pi^x : x \in [0, U]\}$, Assumption 1 is satisfied for all starting states in $\mathcal{S}_{\text{init}}$ (i.e., for $\mathbf{s}_1 \in \mathcal{S}^x_{\text{init}}$). Also, we are given a policy π_{init} such that on using this policy, the expected time to reach a state in $\mathcal{S}^x_{\text{init}}$ from any state using is bounded by D_{init}. Then, the regret of Algorithm 1 is bounded as*

$$Regret(T) = O\Big((H + \beta U)\sqrt{T \log^3(\beta UT/H)} + C_{\max} D_{\text{init}} \log^2(\beta UT/H)\Big).$$

That is,

$$Regret(T) = \tilde{O}((H + \beta U)\sqrt{T} + C_{\max} D_{\text{init}}).$$

Here, $C_{\max}$ is an upper bound on the magnitude of observed costs C^x_t under any policy $\pi^x \in \Pi$; Lipschitz factor β and bias bound H are as defined in Assumption 1.

In this section, we prove the regret bound stated in Theorem 3 for Algorithm 1. Theorem 2 follows as a corollary of Theorem 3, under a stronger condition that Assumption 1 is satisfied for all starting states $\mathbf{s}_1$ so that π_{init} is not required. We first prove a bound on Pseudoregret(T, x) for any $x \in [0, U]$.

The regret analysis follows steps similar to the regret analysis for stochastic convex bandits in Agarwal et al. (2011). We use the notation $f(x) = g^x$ in this proof to connect the regret analysis here to the analysis for stochastic convex bandits with convex function f. Let $x^* := \min_{x\in[0,U]} g^x = \min_{x\in[0,U]} f(x)$. Also, let C_t be the observed cost at time t. Then,

$$\text{Pseudoregret}(T, x) \le \mathbb{E}\left[\sum_{t=1}^{T} C_t\right] - \sum_{t=1}^{T} f(x^*).$$

Also, define $\mathcal{E}$ to be the event that all confidence intervals $[LB(\mathbf{C}^a_N), UB(\mathbf{C}^a_N)]$ calculated in Algorithm 1 satisfy $f(x_a) = g^{x_a} \in [LB(\mathbf{C}^a_N), UB(\mathbf{C}^a_N)]$, where $n = N_i$ for every epoch k, round i, and $a \in \{l, c, r\}$. The analysis in this section will be conditional on event $\mathcal{E}$, and the probability $P(\mathcal{E})$ of this event will be addressed at the end.

We divide the regret into two parts. First, we consider the regret over the set of times steps where policy π_{init} is played. We denote the total contribution of regret from these steps (across all epochs and rounds) as Pseudoregret$^0(T)$. The policy π_{init} is played three times in every round of an epoch. By the assumption stated in Theorem 3, the expected number of steps to reach $\mathcal{S}_{\text{init}}$ is at most D_{init}. Also, the regret in each time step can be at most $C_{\max}$. Therefore, in each round, the expected regret because of the time steps where policy π_{init} is played is bounded by $3D_{\text{init}}C_{\max}$.

Because any epoch has at most T time steps and each successive round within an epoch has four times the number of time steps as the previous (note that $N_{i+1} = 4N_i$), there are at most $\log(T)$ rounds per epoch. Also, in Lemma B.3 (see Appendix B), we show that, under event $\mathcal{E}$, the number of epochs is bounded by $\log_{4/3}\left(\frac{\beta UT}{H}\right)$. Intuitively, this holds because in every epoch, we eliminate at least (1/4)th of the working interval.

Using these observations, the regret from all the time steps where policy π_{init} was executed is bounded by

$$\text{Pseudoregret}^0(T) \le \log_{4/3}\left(\frac{\beta UT}{H}\right)\log(T) \times 3D_{\text{init}}C_{\max}. \tag{10}$$

Next, we consider the regret over all remaining time steps of an epoch, denoted as Pseudoregret$^1(T)$. Algorithm 1 plays the policies π^{x_l}, π^{x_c}, or π^{x_r} in these steps, where x_l, x_c, x_r are updated at the end of every epoch. Consider a round i in epoch k. Let $T_{k,i,l}, T_{k,i,c}, T_{k,i,r}$ be the set of (at most) $N_i = \log(T)/\gamma_i^2$ consecutive time steps where policies $\pi^{x_l}, \pi^{x_c}, \pi^{x_r}$ are played, respectively, in round i of epoch k. Here, $\gamma_i = 2^{-i}$. Let π^{x_t} be the policy used at time t. Note that for $t \in T_{k,i,a}, x_t = x_a$. Given Assumption 1, using Lemma 3 for $\delta = 2/T^2$, for epoch k, round I, and $a \in \{l, c, r\}$, with probability $1 - \delta$,

$$\left|\sum_{t\in T_{k,i,a}} (C_t - f(x_t))\right| \le N_i \times \left(\frac{H}{N_i} + H\sqrt{\frac{2\log(2/\delta)}{N_i}}\right) \le 3H\gamma_i N_i \le 3H\frac{\log(T)}{\gamma_i}.$$

Substituting, we can derive that with probability $1 - 2/T^2$,

$$\begin{aligned}\text{Pseudoregret}^1(T) &= \mathbb{E}\left[\sum_{\text{epoch } k}\ \sum_{\text{round } i}\ \sum_{a\in\{l,c,r\}}\ \sum_{t\in T_{k,i,a}} (C_t - f(x^*))\right] \\ &= \mathbb{E}\left[\sum_{\text{epoch } k}\ \sum_{\text{round } i}\ \sum_{a\in\{l,c,r\}}\ \sum_{t\in T_{k,i,a}} (C_t - f(x_t) + f(x_t) - f(x^*))\right] \\ &\le \mathbb{E}\left[\sum_{\text{epoch } k}\ \sum_{\text{round } i}\left(\frac{9H\log(T)}{\gamma_i} + \sum_{a\in\{l,c,r\}}\ \sum_{t\in T_{k,i,a}} (f(x_t) - f(x^*))\right)\right].\end{aligned} \tag{11}$$

Now, observe that for any round i of epoch k in which the algorithm does not terminate, the total number of time steps is bounded by T. So, for any such k, i, and $a \in \{l, c, r\}$, we have $N_i = \log(T)/\gamma_i^2 \le T$, which implies $\gamma_i \ge \sqrt{\log(T)/T}$. Let us define

$\gamma_{min} := 1/2\sqrt{\log(T)/T}$. Then, because $\gamma_{i+1} = \frac{1}{2}\gamma_i$, we have $\gamma_{min} \le \gamma_j$ for all round j (including the round where the algorithm terminates). Recall that $\gamma_i = 2^{-i}$, so we can bound the geometric series:

$$\sum_k \sum_i \left(\frac{9H\log(T)}{\gamma_i}\right) \le \sum_k \left(\frac{18H\log(T)}{\gamma_{min}}\right) \le \log_{4/3}\left(\frac{\beta UT}{H}\right)\left(36H\sqrt{T\log(T)}\right), \tag{12}$$

where in the last inequality, we used the definition of γ_{min} and the bound $\log_{4/3}(T)$ on the number of epochs from Lemma B.3 (see Appendix B).

Now, consider the second term in (11). We use the results in Agarwal et al. (2011) regarding the convergence of the convex optimization algorithm to bound the gap between $f(x_t)$ and $f(x^*)$. Intuitively, the proof works by showing that the working interval, which contains the optimal solution under event $\mathcal{E}$, shrinks by a constant factor in every epoch, so that $x_t \in \{x_l, x_c, x_r\}$ are closer and closer to the optimal level x^*. Specifically, we adapt the proof from Agarwal et al. (2011) to derive the following bound (see Lemma B.4 in Appendix B) under event $\mathcal{E}$:

$$\sum_{k,i,a,t\in T_{k,i,a}} f(x_t) - f(x^*) \le (12\beta U + 1728H)\log_{4/3}\left(\frac{\beta UT}{H}\right)\sqrt{T\log(T)}. \tag{13}$$

Substituting, in (11), with probability at least $1 - 2/T^2 - \Pr(\neg\mathcal{E})$, Pseudoregret$^1(T)$ is bounded by

$$\text{Pseudoregret}^1(T) \le O((\beta U + H)\log_{4/3}\left(\frac{\beta UT}{H}\right)\sqrt{T\log(T)}. \tag{14}$$

Also, combining with the bound on Pseudoregret$^0(T)$ from (10), we get the following regret bound with probability at least $1 - 2/T^2 - \Pr(\neg\mathcal{E})$:

$$\text{Pseudoregret}(T) \le O\left((\beta U + H)\log_{4/3}\left(\frac{\beta UT}{H}\right)\sqrt{T\log(T)} + C_{\max}D_{\text{init}}\log_{4/3}\left(\frac{\beta UT}{H}\right)\log(T)\right).$$

We complete the proof of the theorem statement by bounding the probability of event $\mathcal{E}$ (i.e., the event that $g^{x_a} \in [LB(\mathbf{C}_N^a), UB(\mathbf{C}_N^a)]$ for every epoch k, round i, $n = N_i$, and $a \in \{l, c, r\}$). By Lemma 3, the condition is satisfied with probability at least $1 - 1/T^2$ for each k, i, a. Because there are no more than T time steps and therefore, at most T plays of any policy, by union bound $\Pr(\mathcal{E}) \ge 1 - 1/T$. Now, using the trivial regret bound of $C_{\max}T$ with probability $3/T$, we get the regret bound, and hence, the derived regret bound holds with probability at least $1 - 1/T$.

Finally, to see that a similar regret bound holds for Regret(T), we compare the two regret definitions:

$$\text{Regret}(T) = \text{Pseudoregret}(T) + Tg^{x^*} - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{x^*} \middle| \mathbf{s}_1\right].$$

Now, using Lemma 3 (for $N = T, \delta = 2/T^2$), the difference $|Tg^{x^*} - \mathbb{E}[\sum_{t=1}^{T} C_t^{x^*}|\mathbf{s}_1]|$ can be bounded by $3H\sqrt{T\log(T)}$ with probability $1 - 2/T^2$. Therefore, we obtain a regret bound on Regret(T) of the same order as Pseudoregret(T).

3. Improved Regret Bounds for Inventory Management

In this section, we apply the algorithm and regret analysis presented in the previous section to the inventory control problem. Specifically, we establish that the convexity and bounded bias properties stated in Assumption 1 hold for the MRP obtained on running a base-stock policy for the inventory control problem, so that Theorem 3 can be applied to obtain our main result stated in Theorem 1.

To define the MRP studied here, we observe that if we start with an on-hand inventory level and a pipeline of outstanding orders that sum to less than or equal to x, then on using the base-stock policy with level x, the new order o_t will bring the sum to exactly x: that is,

$$o_t = x - \left(\text{inv}_t + \sum_{i=1}^{L} o_{t-i}\right) = x - I_t - \sum_{i=1}^{L-1} o_{t-i}$$

because $I_t = \text{inv}_t + o_{t-L}$. From here on, the base-stock policy will always order whatever is consumed because of demand (i.e., $o_{t+1} = y_t$, where $y_t = \min\{I_t, d_t\}$ is the observed sales), so that the sum of on-hand inventory level and outstanding orders will be maintained at level x.

3.1. MRP Formulation

We define an MRP with state $\mathbf{s}_t$ at time t defined as the tuple of available inventory and outstanding orders (including the new order) (i.e., $\mathbf{s}_t = (I_t, o_{t-L+1}, \ldots, o_t)$). The MRP starts from a state $\mathbf{s}_1$ such that the inventory position (i.e., the sum of on-hand inventory level and outstanding orders) is equal to x. Then, based on the observation made, the base-stock policy will maintain the inventory position at level x, with the new state at time $t+1$ being $\mathbf{s}_{t+1} = (I_t - y_t + o_{t-L+1}, o_{t-L+2}, \ldots, o_t, o_{t+1})$, where $o_{t+1} = y_t$. We define the cost associated with each state $\mathbf{s}_t$ in this MRP as $C^x(\mathbf{s}_t) = \mathbb{E}[C_t^x|\mathbf{s}_t]$, where C_t^x is defined as the following modification of the true cost $\bar{C}_t^x$:

$$C_t^x = \bar{C}_t^x - pd_t = h(I_t - y_t) - py_t. \tag{15}$$

The advantage of using this modified cost is that because both I_t (on-hand inventory) and y_t (sales) are observable, the pseudocost is completely observed. On the other hand, recall that the "lost sales" in the true cost are not observed. Further, because the term pd_t does not depend on the policy or the algorithm being used, for any two policies $\pi^x, \pi^{x'}$, $C_t^x - C_t^{x'} = \bar{C}_t^x - \bar{C}_t^{x'}$. Therefore, the regret of an algorithm under the modified cost is the same as the regret under the true cost; from here on, we use the regret definition under the modified cost,

$$\text{Regret}(T,x) = \mathbb{E}\left[\sum_{t=1}^{T} C_t \middle| \mathbf{s}_1\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^* \middle| \mathbf{s}_1\right]. \tag{16}$$

Definition 4 is the precise definition of the state space, starting state, reward model, and transition model of the MRP considered here.

Definition 4 (Markov Reward Process $\mathcal{M}(\boldsymbol{x}, \mathbf{s}_1)$). For any $x \geq 0$, let $\mathcal{S}_{\text{init}}^x$ be the set of $(L+1)$-dimensional nonnegative vectors whose components sum to x. Then, given any x and $\mathbf{s}_1 \in \mathcal{S}_{\text{init}}^x$, we define MRP $\mathcal{M}(x, \mathbf{s}_1)$ as the bipartite stochastic process $\{(\mathbf{s}_t, C^x(\mathbf{s}_t)); t = 1, 2, 3, \dots\}$.

Here, $\mathbf{s}_t$ and $C^x(\mathbf{s}_t)$ denote the state and the cost at time t, defined as follows. Given state $\mathbf{s}_t = (s_t(0), s_t(1), \dots, s_t(L))$, the new state at time $t+1$ is given by

$$\mathbf{s}_{t+1} := (s_t(0) - y_t + s_t(1), s_t(2), \dots, s_t(L), y_t),$$

where $y_t = \min\{s_t(0), d_t\}, d_t \sim F$, generated independently from distribution F at every time t. Observe that if $\mathbf{s}_1 \in \mathcal{S}_{\text{init}}^x$, we have $\mathbf{s}_t \in \mathcal{S}_{\text{init}}^x$ for all t by the transition process. That is, $\mathcal{S}^x = \mathcal{S}_{\text{init}}^x$, the set of states where all components sum to x. Cost function $C^x(\mathbf{s}_t)$ is defined as

$$C^x(\mathbf{s}_t) = \mathbb{E}[C_t^x | \mathbf{s}_t],$$

where

$$C_t^x := h(s_t(0) - y_t) - py_t. \tag{17}$$

Two important quantities are the *loss* $g^x(\mathbf{s})$ and *bias* $v^x(\mathbf{s})$ of this MRP for any state $\mathbf{s} \in \mathcal{S}^x$, which are as defined in Definition 3:

$$g^x(\mathbf{s}) := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} C^x(\mathbf{s}_t) \middle| \mathbf{s}_1 = \mathbf{s}\right]$$
$$v^x(\mathbf{s}) := \mathbb{E}\left[\lim_{T\to\infty} \sum_{t=1}^{T} C^x(\mathbf{s}_t) - g^x(\mathbf{s}_t) \middle| \mathbf{s}_1 = \mathbf{s}\right].$$

Remark 3. Technically, for the limits to exist and also, for some other known results on MRPs used later, we need finite state space and finite action space (see chapter 8.2 in Puterman 2014). Because we restrict to orders within range $[0, U]$ and all states $\mathbf{s} \in \mathcal{S}^x$ are vectors in $[0, x]^L$ with $x \in [0, U]$, we can obtain finite state space and action space by discretizing demand and orders using a uniform grid with spacing $\epsilon \in (0, 1)$. Discretizing this way will give us a state space and action space of size $(U/\epsilon)^L$ and U/ϵ, respectively. In fact, we can use arbitrary small precision parameter ϵ because our bounds will not depend on the size of the state space or the action space. We, therefore, ignore this technicality in rest of the paper.

Next, we show that conditions stated in Assumption 1 are satisfied by this MRP and derive a bound H on bias. To prove these results, we find it useful to consider another related quantity, called the value of a policy in finite time T.

Definition 5 (Value). For any $\mathbf{s} \in \mathcal{S}^x$, the value $V_T^x(\mathbf{s})$ is defined as the total expected cost incurred over T time steps of MRP $M(x, \mathbf{s})$: that is,

$$V_T^x(\mathbf{s}) := \mathbb{E}\left[\sum_{t=1}^{T} C^x(\mathbf{s}_t) \middle| \mathbf{s}_1 = \mathbf{s}\right].$$

We first prove a bound on the difference in value of two states under any given base-stock policy and then, use that bound to prove the properties of (a) uniform loss, (b) convexity of loss, and (c) bounded span of bias, as needed for Assumption 1.

3.2. Bounded Difference in Value

Lemma 5 (Bounded Difference in Value). *For any x, T, and $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$,*

$$V_T^x(\mathbf{s}) - V_T^x(\mathbf{s}') \leq 36\max(h, p)Lx.$$

Proof. For $L = 0$, $\mathbf{s} = \mathbf{s}' = (x)$, and hence, both sides are zero in the inequality. Consider $L \geq 1$. One way to bound the difference between the two values $V_T^x(\mathbf{s})$ and $V_T^x(\mathbf{s}')$ is to upper bound the expected number of steps for the MRPs to reach a common state, starting from $\mathbf{s}$ and $\mathbf{s}'$. Once a common state is reached, from that point onward, the two processes will have the same value. For example, in the event that there is zero demand for L consecutive time steps, both processes will order 0 for L time steps and reach state $(x, 0, \dots, 0)$. Therefore, the difference in values can be upper bounded by the expected number of steps until this event happens, which is a quantity proportional to the inverse of the probability that demand is zero for L consecutive steps. Unfortunately, this probability is exponentially small in L. In fact, the exponential dependence of regret in previous works (Huh et al. 2009a, Zhang et al. 2020) can be traced to use of an argument like the one given at some point in the analysis. Instead, we achieve a bound with linear dependence on L through a more careful analysis of the costs incurred on starting from different states.

For any $\mathbf{s} \in \mathcal{S}^x$, we define $m_T^x(\mathbf{s}) := \sum_{t=1}^{T} s_t(0)$ to be the total on-hand inventory level (recall $s_t(0) = I_t$) and

$n_T^x(\mathbf{s}) := \sum_{t=1}^T y_t$ to be the total sales in T time steps (recall $y_t = \min(s_t(0), d_t), d_t \sim F$) on starting from state $\mathbf{s}$. Then,

$$V_T^x(\mathbf{s}) := \mathbb{E}\left[\sum_{i=1}^T C_t^x \middle| \mathbf{s}_1 = \mathbf{s}\right] = \mathbb{E}\left[\sum_{i=1}^T hI_t - (h+p)y_t \middle| \mathbf{s}_1 = \mathbf{s}\right]$$
$$= h\mathbb{E}[m_T^x(\mathbf{s})] - (h+p)\mathbb{E}[n_T^x(\mathbf{s})].$$

Thus, the difference between values $V_T^x(\mathbf{s})$ and $V_T^x(\mathbf{s}')$ can be bounded by bounding the difference in the total on-hand inventory $|m_T^x(\mathbf{s}) - m_T^x(\mathbf{s}')|$ and the total sales $|n_T^x(\mathbf{s}) - n_T^x(\mathbf{s}')|$ when starting from state $\mathbf{s}'$ versus $\mathbf{s}$. We bound this difference by first considering pairs of states $\mathbf{s}, \mathbf{s}'$ that satisfy $\mathbf{s}' \succeq \mathbf{s}$, with the relation $\succeq$ defined as the property that for some index $k \geq 0$, the first k entries satisfy $s'(0) \geq s(0), \ldots, s'(k) \geq s(k)$ and the remaining $L+1-k$ entries satisfy $s'(k+1) \leq s(k+1), \ldots, s'(L) \leq s(L)$.

When $\mathbf{s}' \succeq \mathbf{s}$ and $L \geq 1$, we bound the difference in the total sales and the total on-hand inventory as

$$|n_T^x(\mathbf{s}) - n_T^x(\mathbf{s}')| \leq 3x \text{ and } |m_T^x(\mathbf{s}) - m_T^x(\mathbf{s}')| \leq 6Lx. \quad (18)$$

To see the intuition behind proving these bounds, consider the sales observed on starting from $\mathbf{s}'$ versus $\mathbf{s}$. Recall that the first entry in $\mathbf{s}$ (and $\mathbf{s}'$) is the on-hand inventory, the second entry is the order to arrive next, the next entry is the order to arrive after that, and so on. Therefore, intuitively, $\mathbf{s}' \succeq \mathbf{s}$ implies that initially more inventory is available on hand to satisfy demand, when starting from $\mathbf{s}'$. We use this intuition to show that indeed more sales are observed initially on starting from $\mathbf{s}'$ compared with $\mathbf{s}$. Over time, the two processes keep alternating between states with $\mathbf{s}_t' \succeq \mathbf{s}_t$ and $\mathbf{s}_t' \preceq \mathbf{s}_t$ in cycles of length at most L. The additional sales in one cycle with $\mathbf{s}_t' \succeq \mathbf{s}_t$ compensate for the lower sales in the next cycle with $\mathbf{s}_t' \preceq \mathbf{s}_t$, so that the total difference is bounded. We also remark that the linear dependence on L can be traced to the upper bound on the difference in total on-hand inventory between two different states in (18). This bound is tight (with respect to L) for the case when $\mathbf{s} = (x, 0, \ldots, 0)$, $\mathbf{s}' = (0, \ldots, 0, x)$, and there is zero demand for L consecutive time steps. The formal proofs for bounding the difference in total sales and on-hand inventory for the case $\mathbf{s}' \succeq \mathbf{s}$ are provided in Lemmas C.4 and C.5, respectively, in the appendix.

Then, we use the observation that $\hat{\mathbf{s}} \succeq \mathbf{s}$ for all states $\mathbf{s} \in \mathcal{S}^x$ for $\hat{\mathbf{s}} := (x, 0, 0, \ldots, 0)$. Therefore, we can apply the result in (18) to conclude that

$$|n_T^x(\mathbf{s}) - n_T^x(\hat{\mathbf{s}})| \leq 3x \text{ and } |m_T^x(\mathbf{s}) - m_T^x(\hat{\mathbf{s}})| \leq 6Lx,$$

implying

$$|V_T^x(\mathbf{s}) - V_T^x(\hat{\mathbf{s}})| = |h\mathbb{E}[m_T^x(\mathbf{s}) - m_T^x(\hat{\mathbf{s}})] - (h+p)\mathbb{E}[n_T^x(\mathbf{s}) - n_T^x(\hat{\mathbf{s}})]| \leq 9(h+p)Lx.$$

Because this holds for any state $\mathbf{s}$, we have that for two arbitrary starting states $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$,

$$|V_T^x(\mathbf{s}) - V_T^x(\mathbf{s}')| = |V_T^x(\mathbf{s}) - V_T^x(\hat{\mathbf{s}}) + V_T^x(\hat{\mathbf{s}}) - V_T^x(\mathbf{s}')| \leq 18(h+p)Lx \leq 36\max(h,p)Lx. \quad \square$$

3.3. Uniform and Convex Loss

Next, we use the value difference lemma (Lemma 5) to show that the loss $g^x(\mathbf{s})$ (refer to Definition 3) is independent of the starting state $\mathbf{s} \in \mathcal{S}^x$ in the inventory control MRP.

Lemma 6 (Uniform Loss Lemma). *For any x, $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$,*

$$g^x(\mathbf{s}') = g^x(\mathbf{s}) =: g^x.$$

Proof. Using the definition of $V_T^x(\mathbf{s})$ and $g^x(\mathbf{s})$, $g^x(\mathbf{s}) = \lim_{T\to\infty} 1/T V_T^x(\mathbf{s})$ so that by Lemma 5,

$$|g^x(\mathbf{s}) - g^x(\mathbf{s}')| = \left|\lim_{T\to\infty} \frac{1}{T} V_T^x(\mathbf{s}) - \lim_{T\to\infty} \frac{1}{T} V_T^x(\mathbf{s}')\right| \leq \lim_{T\to\infty} \frac{36\max(h,p)Lx}{T} = 0$$

because both limits exist (see Remark 3). Hence, for any $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$, $g^x(\mathbf{s}') = g^x(\mathbf{s})$.

Now, the convexity of g^x follows almost immediately from convexity of long-run average cost λ^x for any base-stock policy, proven by Janakiraman and Roundy (2004) (refer to Lemma 1).

Lemma 7 (Convexity Lemma). *Assume that demand distribution F is such that there is a constant probability of zero demand (i.e., $F(0) > 0$). Then, for any base-stock level x and $\mathbf{s} \in \mathcal{S}^x$, $g^x(\mathbf{s})$ is convex in x.*

Proof. Let $\mathbf{s}' := (x, 0, \ldots, 0)$, and let μ be the mean of the demand distribution F. On using the base-stock policy with level x, starting with no inventory and no outstanding orders, the first order will be x, which will arrive at time step $L+1$. The orders and the on-hand inventory will be zero for the first L time steps ($I_1 = I_2 = \ldots = I_L = 0$). All the sales will be lost for these first L steps, and therefore, the true cost $\bar{C}_t^x$ in each of these steps is the lost sales penalty pd_t. In the step $L+1$, we will have an on-hand inventory $I_{L+1} = x$ and no outstanding orders. From here on, the system will follow a Markov reward process $\mathcal{M}(x, \mathbf{s})$ with $\mathbf{s}_1 = \mathbf{s}'$. Therefore, by the relation (see (15)) between the modified cost C_t^x and the true cost (lost sales penalty plus holding cost) $\bar{C}_t^x$, we have $C_t^x = \bar{C}_t^x - pd_t$, for $t \geq L+1$. Therefore,

$$g^x(\mathbf{s}') = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=L+1}^{L+T} C^x(\mathbf{s}_t) \middle| \mathbf{s}_{L+1} = \mathbf{s}'\right]$$
$$= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=L+1}^{T+L} (\bar{C}_t^x - pd_t) \middle| I_{t+L} = x, o_1 = o_2 = \cdots = o_L = 0\right]$$

$$= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T+L}(\bar{C}_t^x - pd_t)\,\middle|\,\text{inv}_1 = 0\right] = \lambda^x - p\mu.$$

Therefore, under the given assumption that the demand distribution F has a nonzero probability of zero demand, we can use Lemma 1 to conclude that the first term is convex in x, which implies that $g^x(\mathbf{s}')$ is convex. Now, by Lemma 6, for any state $\mathbf{s} \in \mathcal{S}^x$, $g^x(\mathbf{s}) = g^x(\mathbf{s}')$. Therefore, $g^x(\mathbf{s})$ is convex in x for all $\mathbf{s} \in \mathcal{S}^x$.

We also prove the following bound on the Lipschitz factor of g^x.

Lemma 8 (Lipschitz Factor of g^x). *The function g^x is Lipschitz continuous in x for $x \in [0, U]$, with the Lipschitz factor of $\beta = \max(h, p)$. That is, for any $\delta \geq 0$,*

$$|g^{x+\delta} - g^x| \leq \max(h,p)\delta.$$

Proof. Let us compare the loss $g^{x+\delta}$ versus g^x on executing the base-stock policy with level $x+\delta$ versus x. Let us assume that the starting states for the two MRPs are $\mathbf{s}_1^1 = (x+\delta, 0, \ldots, 0) \in \mathcal{S}^{x+\delta}$ and $\mathbf{s}_1^2 = (x, 0, \ldots, 0) \in \mathcal{S}^x$, respectively (this is without loss of generality because recall from Lemma 6 that loss is independent of the starting state). We compare the two losses by coupling the execution of the two MRPs. For every time t, let $\mathbf{s}_t^1 := (I_t^1, o_{t-L+1}^1, \ldots, o_t^1)$ be the state of the system on following the base-stock policy with level $x+\delta$ and $\mathbf{s}_t^2 := (I_t^2, o_{t-L+1}^2, \ldots, o_t^2)$ be the state of the system on following the policy with level x. Define $\mathbf{s}_1^1 \geq \mathbf{s}_1^2$ if every entry in $\mathbf{s}_1$ is at least the respective entry in $\mathbf{s}_2$.

We will first show by induction that at each time step t, $\mathbf{s}_t^1 \geq \mathbf{s}_t^2$. In the first time step, we have $\mathbf{s}_1^1 = (x+\delta, 0, \ldots, 0) \geq (x, 0, \ldots, 0) = \mathbf{s}_1^2$. From then on, the new order placed at time $t+1$ is the amount of sales in the previous time step t. Therefore, if at time t, we have that $\mathbf{s}_t^1 \geq \mathbf{s}_t^2$, then the orders at time $t+1$ satisfy $o_{t+1}^1 = \min(d_t, I_t^1) \geq \min(d_t, I_t^2) = o_{t+1}^2$. Also, $I_{t+1}^1 = (I_t^1 - \min(d_t, I_t^1)) + o_{t-L}^1 \geq (I_t^2 - \min(d_t, I_t^2)) + o_{t-L}^2 = I_{t+1}^2$. Hence, we have $\mathbf{s}_{t+1}^1 \geq \mathbf{s}_{t+1}^2$. By induction, we have that for every $t \geq 1$, $\mathbf{s}_t^1 \geq \mathbf{s}_t^2$.

We complete the proof by noting that, additionally, at every time t, the total sum of the entries of $\mathbf{s}_t^1$ is exactly δ greater than the sum of the entries of $\mathbf{s}_t^2$. Therefore, the difference $0 \leq I_t^1 - I_t^2 \leq \delta$ for every t, which implies the difference in sales $0 \leq y_t^1 - y_t^2 = \min(d_t, I_t^1) - \min(d_t, I_t^2) \leq \delta$. This implies $0 \leq (I_t^1 - y_t^1) - (I_t^2 - y_t^2) \leq \delta$. Recall pseudocost $C_t^{x+\delta} = (I_t^1 - y_t^1)h - py_t^1$ and $C_t^x = (I_t^2 - y_t^2)h - py_t^2$; therefore, we have that for every t and every sequence of demand realizations,

$$|C_t^{x+\delta} - C_t^x| \leq \max(h,p)\delta.$$

By definition of loss g^x as the long-run average of these costs (see Definition 3), we have

$$|g^{x+\delta} - g^x| \leq \max(h,p)\delta.$$

3.4. Bound on Bias

Our main technical insight for the inventory control problem is the following bound on bias under any base-stock policy.

Lemma 9 (Bounded Bias Lemma). *For any x and $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^x$, the difference in bias $v^x(\mathbf{s}) - v^x(\mathbf{s}')$ as*

$$v^x(\mathbf{s}) - v^x(\mathbf{s}') \leq 36\max(h,p)Lx.$$

That is, for all $x \in [0, U]$, we have that the span of bias is bounded by $H = 36\max(h,p)LU$.

Proof. From Lemma 6, $g^x(\mathbf{s}_t) = g^x(\mathbf{s}_t') = g^x$ for all t. Now, by definition of bias $v^x(\cdot)$ (refer to Definition 3),

$$v^x(\mathbf{s}) = \mathbb{E}\left[\lim_{T\to\infty}\sum_{t=1}^{T} C^x(\mathbf{s}_t) - g^x \,\middle|\, \mathbf{s}_1 = \mathbf{s}\right] = \lim_{T\to\infty} V_T^x(\mathbf{s}) - Tg^x$$

and

$$\begin{aligned} v^x(\mathbf{s}') &= \mathbb{E}\left[\lim_{T\to\infty}\sum_{t=1}^{T} C^x(\mathbf{s}_t) - g^x \,\middle|\, \mathbf{s}_1 = \mathbf{s}'\right] \\ &= \lim_{T\to\infty} V_T^x(\mathbf{s}') - Tg^x. \end{aligned}$$

We note that both of the limits exist (see Remark 3), and hence, by Lemma 5,

$$\begin{aligned} v^x(\mathbf{s}) - v^x(\mathbf{s}') &= \lim_{T\to\infty}(V_T^x(\mathbf{s}) - Tg^x) - \lim_{T\to\infty}(V_T^x(\mathbf{s}') - Tg^x) \\ &= \lim_{T\to\infty} V_T^x(\mathbf{s}) - V_T^x(\mathbf{s}') \\ &\leq 36\max(h,p)Lx. \end{aligned}$$

3.5. Algorithm and Regret Bound

We apply Algorithm 1 to the inventory control problem with the following inputs.

- Π is the set of base-stock policies π^x with $x \in [0, U]$.
- $H = 36\max(h,p)LU$.
- $\mathcal{S}_{\text{init}}^x = \mathcal{S}^x$ is the set of all vectors $\mathbf{s} = (s(0), s(1), \ldots, s(L))$ such that $\mathbf{s} \geq 0$ and $\sum_{i=0}^{L} s(i) = x$.
- π_{init}^x is the policy of ordering zero until $\sum_{i=0}^{L} s(i) \leq x$ and then, using policy π^x once so that $\sum_{i=0}^{L} s(i) = x$.

Then, we use the convexity property and bound on bias proven in the previous section to obtain the regret bound stated in Theorem 1.

Proof of Theorem 1. By Lemmas 6–9, the conditions (a)–(c) of Assumption 1 are satisfied with $H = 36LU\max(h,p)$ and $\beta = \max(h,p)$.

Further, we observe that D_{init}, the expected time to reach the set of states $\mathcal{S}_{\text{init}}$ defined, is bounded by $L + DU$ under the policy π_{init} defined. To see this, note that under this policy, an order of zero units is placed until inventory position falls below x. Note that ordering zero for L steps will result in at most U inventory on hand and no orders in the pipeline. By definition of D, the expected number of time steps to deplete U

units of inventory is upper bounded by DU. Therefore, the expected time for inventory position to fall below x is at most $L+DU$. Then, using policy π^x once will bring the inventory position to x (i.e., the state to be in $\mathcal{S}^x_{\text{init}}$). Therefore, $D_{\text{init}} \le L+DU+1$.

Also, $C^t_x = h(I_t - y_t) - py_t \le \max(h,p)(2|y_t| + |I_t|) \le 3\max(h,p)U$. Therefore, $C_{\max} = 3\max(h,p)U$ upper bounds the observed costs.

Substituting $C_{\max} = 3U\max(h,p), H = 36LU\max(h,p)$, $\beta = \max(h,p)D_{\text{init}} = L+DU+1$ in Theorem 3, we obtain the result stated in Theorem 1. □

4. Conclusions

We presented an exploration-exploitation algorithm to minimize regret in the periodic inventory control problem under censored demand, lost sales, and positive lead time when compared with the best base-stock policy. By using convexity properties of the long-run average cost function and a newly proven bound on the bias of base-stock policies, we extended a stochastic convex bandit algorithm to obtain a simple algorithm that substantially improves upon the existing solutions for this problem. In particular, the regret bound for our algorithm maintains an optimal dependence on T while also achieving a linear dependence on lead time. The algorithm design and analysis techniques were presented for any general MDP satisfying the convexity and bounded bias properties. We believe that these techniques could be useful for obtaining efficient solutions for other classes of learning problems where the MDPs involved may be large, but the long-run average cost under benchmark policies is known to be convex.

Acknowledgments

The authors thank the anonymous associate editor and reviewers for their valuable comments. A preliminary version of this paper was accepted to the 20th Association for Computing Machinery (ACM) Conference on Economics and Computation 2019.

Appendix A. Concentration Bounds

Lemma A.1 (Azuma–Hoeffding Inequality). *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \le c$ for all i. Then, for all $\epsilon > 0$ and $n \in \mathbb{N}$,*

$$P\left(\left|\sum_{i=1}^{n} X_i\right| \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2}{2nc^2}\right).$$

Appendix B. Proof Details for Theorem 3

We present additional results required to complete the proof of Theorem 3. Recall that $f(x) := g^x$ is a convex function. Also, given confidence intervals defined as in (9), recall that $\mathcal{E}$ is the event when all confidence intervals $[LB(C^a_N), UB(C^a_N)]$ calculated in all rounds of Algorithm 1 satisfy $g^{x_a} \in [LB(C^a_N), UB(C^a_N)]$ for $N \ge N_i$ in every round i, every epoch k, and $a \in \{l, c, r\}$.

The proofs for the lemmas provided are similar to the proofs of the corresponding lemmas in Agarwal et al. (2011). We include the proofs here for completeness.

Lemma B.1 (Lemma 1 in Agarwal et al. 2011). *Recall $[l_k, r_k]$ denotes the working interval in epoch k of Algorithm 1, with $[l_1, r_1] := [0, U]$. Then, under event $\mathcal{E}$, for epoch k ending in round i, the working interval $[l_{k+1}, r_{k+1}]$ for the next epoch $k+1$ contains every $x \in [l_k, r_k]$ such that $f(x) \le f(x^*) + 6H\gamma_i$. In particular, $x^* \in [l_k, r_k]$ for all epochs k.*

Proof. Consider any epoch k that is not the last epoch. Then, under Algorithm 1, if the epoch k ends in round i, then

$$\max\{LB(C^l_N), LB(C^r_N)\} \ge \min\{UB(C^l_N), UB(C^c_N), UB(C^r_N)\} + 6H\gamma_i,$$

where $N = N_i$. Hence,

1. $LB(C^l_N) \ge UB(C^r_N) + 6H\gamma_i$,
2. $LB(C^r_N) \ge UB(C^l_N) + 6H\gamma_i$, or
3. $\max\{LB(C^l_N), LB(C^r_N)\} \ge UB(C^c_N) + 6H\gamma_i$.

Consider case (1) (case (2) is analogous). Then,

$$f(x_l) \ge f(x_r) + 6H\gamma_i. \tag{B.1}$$

We need to show that every $x \in [l_k, l_{k+1}]$ has $f(x) \ge f(x^*) + 6H\gamma_i$ (for this case, note that $r_{k+1} = r_k$, and hence, the interval $[r_{k+1}, r_k]$ is of length zero). Pick $x \in [l_k, x_l]$ so that $x_l \in [x, x_r]$. Then, $x_l = \lambda x + (1-\lambda)x_r$ for some $0 \le \lambda \le 1$ so by convexity,

$$f(x_l) \le \lambda f(x) + (1-\lambda) f(x_r).$$

This implies that

$$f(x) \ge f(x_r) + \frac{f(x_l) - f(x_r)}{\lambda} \ge f(x_r) + \frac{6H\gamma_i}{\lambda} \ge f(x^*) + 6H\gamma_i,$$

where we used (B.1) and that $\lambda \le 1$.

Now, consider case (3). Assume $LB(C^l_N) \ge LB(C^r_N)$ (the other case is analogous). Then, in case (3), we have

$$f(x_l) \ge f(x_c) + 6H\gamma_i.$$

We need to show that every $x \in [l_k, l_{k+1}]$ has $f(x) \ge f(x^*) + 6H\gamma_i$. (Again, $r_{k+1} = r_k$, and hence, the interval $[r_{k+1}, r_k]$ is of length zero.) This follows from the same argument with x_r replaced by x_c. The fact that $x^* \in [l_k, r_k]$ for all epochs k follows by induction. □

Lemma B.2 (Lemma 2 in Agarwal et al. 2011). *Under $\mathcal{E}$, if epoch k does not end in round i, then $f(x) \le f(x^*) + 72H\gamma_i$ for each $x \in \{x_r, x_c, x_l\}$.*

Proof. Under Algorithm 1, round i continues to round $i+1$ if

$$\max\{LB(C^l_N), LB(C^r_N)\} < \min\{UB(C^l_N), UB(C^c_N), UB(C^r_N)\} + 6H\gamma_i.$$

We observe that because each confidence interval is of length $6H\gamma_i$, this means that $f(x_l), f(x_c), f(x_r)$ are contained within an interval of length at most $18H\gamma_i$. By Lemma B.1, $x^* \in [l_k, r_k]$. Under event $\mathcal{E}$, without loss of generality, assume $x^* \le x_c$ (the other case is analogous). Then, there exists $\lambda \ge 0$ such that $x^* = x_c + \lambda(x_c - x_r)$, so that

$$x_c = \frac{1}{1+\lambda} x^* + \frac{\lambda}{1+\lambda} x_r.$$

Note that $\lambda \le 2$; this is because $|x_c - l_k| = \frac{w_k}{2}$ and $|x_r - x_c| = w_k/4$, so that

$$\lambda = \frac{|x^* - x_c|}{|x_r - x_c|} \le \frac{|l_k - x_c|}{|x_r - x_c|} = \frac{w_k/2}{w_k/4} = 2.$$

Now, because f is convex,

$$f(x_c) \le \frac{1}{1+\lambda} f(x^*) + \frac{\lambda}{1+\lambda} f(x_r),$$

and so,

$$\begin{aligned} f(x^*) &\ge (1+\lambda)\left(f(x_c) - \frac{\lambda}{1+\lambda} f(x_r)\right) \\ &= f(x_r) + (1+\lambda)(f(x_c) - f(x_r)) \\ &\ge f(x_r) - (1+\lambda)|f(x_c) - f(x_r)| \\ &\ge f(x_r) - (1+\lambda)18H\gamma_i \\ &\ge f(x_r) - 54H\gamma_i. \end{aligned}$$

Thus, for each $x \in \{x_l, x_c, x_r\}$,

$$f(x) \le f(x_r) + 18H\gamma_i \le f(x^*) + 72H\gamma_i.$$

Lemma B.3 (Lemma 4 in Agarwal et al. 2011). *Under $\mathcal{E}$, the total number of epochs K is bounded by $\log_{4/3}(T)$.*

Proof. Observe that for any round i that does not terminate the algorithm, $N_i = \log(T)/\gamma_i^2 \le T$ (because the algorithm terminates upon reaching T time steps), which implies $\gamma_i \ge \sqrt{\log(T)/T}$. Because $\gamma_{i+1} = (1/2)\gamma_i$, let us define $\gamma_{min} := (1/2)\sqrt{\log(T)/T}$ so that $\gamma_{min} \le \gamma_i$ for any γ_i. Define the interval $I := [x^* - 6H\gamma_{\min}/\beta, x^* + 6H\gamma_{\min}/\beta]$, so that for any $x \in I$,

$$f(x) - f(x^*) \le \beta|x - x^*| \le 6H\gamma_{\min}$$

because the Lipschitz factor of f is β. Now, for any epoch k', which ends in round i', $H\gamma_{\min} \le H\gamma_{i'}$, and hence, by Lemma B.1, we have

$$I \subseteq \{x \in [0,U] : f(x) \le f(x^*) + 6H\gamma_{i'}\} \subseteq [l_{k'+1}, r_{k'+1}].$$

So, for any epoch k', the length of interval I is no more than the length of interval $[l_{k'+1}, r_{k'+1}]$, and so,

$$\frac{12H\gamma_{\min}}{\beta} \le r_{k'+1} - l_{k'+1} =: w_{k'+1}.$$

Because $w_{k'+1} \le 3/4 w_{k'}$ for any $k' = 1, 2, \ldots, K-1$, we have that for $k' = K-1$,

$$\frac{12H\gamma_{\min}}{\beta} = \frac{6H}{\beta}\sqrt{\frac{\log(T)}{T}} \le w_K \le \left(\frac{3}{4}\right)^{K-1} w_1 = \left(\frac{3}{4}\right)\left(\frac{3}{4}\right)^K (U).$$

Rearranging the inequality, we get that

$$K \le \frac{1}{2}\log_{4/3}\left(\frac{9\beta^2(U)^2 T}{16(6H)^2\log(T)}\right) \le \log_{4/3}\left(\frac{\beta UT}{H}\right).$$

Lemma B.4 (Lemma 3 in Agarwal et al. 2011). *Recall $T_{k,i,a}$ is the set of consecutive time steps where base-stock policy level x_a is played in round i of epoch k, for $a \in \{l, c, r\}$. Then, under $\mathcal{E}$, we can bound*

$$\begin{aligned} \sum_{t=1}^{T} f(x_t) - f(x^*) &= \sum_{k,i,a,t \in T_{k,i,a}} f(x_t) - f(x^*) \\ &\le (12\beta U + 1728H)\log_{4/3}\left(\frac{\beta UT}{H}\right)\sqrt{T\log(T)}. \end{aligned}$$

Proof. Let us first fix an epoch k and assume it ends in round $i(k)$. If $i(k) = 1$, then

$$\sum_{i,a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) \le 3N_1\beta|x_t - x^*| \le \left(\frac{3\log(T)}{\gamma_1^2}\right)\beta U. \quad \text{(B.2)}$$

Otherwise, if $i(k) > 1$, then

$$\begin{aligned} \sum_{i,a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) &= \sum_{i=1}^{i(k)-1} \sum_{a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) \\ &\quad + \sum_{a,t \in T_{k,i(k),a}} (f(x_t) - f(x^*)). \end{aligned}$$

By Lemma B.2, for each $x_t \in \{x_r, x_c, x_l\}$, $f(x_t) - f(x^*) \le 72H\gamma_i$ for all $i = 1, 2, \ldots, i(k) - 1$. Also, $\gamma_{i(k)-1} = 2\gamma_{i(k)}$, so when $i(k) > 1$,

$$\begin{aligned} \sum_{i,a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) &\le \sum_{i=1}^{i(k)-1} \sum_{a,t \in T_{k,i,a}} (72H\gamma_i) + \sum_{a,t \in T_{k,i(k),a}} (72H\gamma_{i(k)-1}) \\ &\le \sum_{i=1}^{i(k)-1} (3N_i)(72H\gamma_i) + (3N_{i(k)})(144H\gamma_{i(k)}) \\ &\le \sum_{i=1}^{i(k)} 432 N_i H\gamma_i \\ &\le \frac{864H\log(T)}{\gamma_{min}}, \end{aligned}$$

where in the last step, we used that

$$\sum_i \frac{1}{\gamma_i} \le \frac{1}{\gamma_{min}}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \ldots\right) = \frac{2}{\gamma_{min}}.$$

Combining this result with (B.2), we have for any epoch k, irrespective of number of rounds $i(k)$,

$$\begin{aligned} \sum_{i,a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) &\le \left(\frac{3\log(T)}{\gamma_1^2}\right)\beta U + \frac{864H\log(T)}{\gamma_{min}} \\ \text{(because } \gamma_{min} \le \gamma_1 = 1/2) &\le \left(\frac{6\log(T)}{\gamma_{min}}\right)\beta U + \frac{864H\log(T)}{\gamma_{min}} \\ &\le \frac{\log(T)}{\gamma_{min}}(6\beta U + 864H). \end{aligned}$$

Therefore, summing over all epochs k, by Lemma B.3,

$$\sum_{k,i,a,t \in T_{k,i,a}} (f(x_t) - f(x^*)) \le \log_{4/3}\left(\frac{\beta UT}{H}\right)\frac{\log(T)}{\gamma_{min}}(6\beta U + 864H),$$

and the result follows from substituting $\gamma_{min} = (1/2)\sqrt{\log(T)/T}$. □

Appendix C. Proof Details for Lemma 5

In this section, we provide the proof details for results used in Lemma 5 when $L \ge 1$. Recall that MRP $\mathcal{M}(x, \mathbf{s}_1)$ is defined such that state $\mathbf{s} = (s(0), s(1), \ldots, s(L))$, with $s(0)$ being the on-hand inventory after the current time step's order arrival and new order, and $s(1), \ldots, s(L)$ are outstanding orders, with $s(L)$ being the most recent order scheduled to arrive L time steps after the current time. The process starts with a state $\mathbf{s}_1 \in \mathcal{S}^x$ (i.e., a state $\mathbf{s}_1$ such that $s_1(0) + \ldots + s_1(L) = x$). Then,

because of the use of the base-stock policy with level x, new orders are placed such that at every time step $t = 1, 2, \ldots, T$, we have $\sum_{i=0}^{L} \mathbf{s}_t(i) = x$ (i.e., $\mathbf{s}_t \in \mathcal{S}^x$), where $\mathbf{s}_t$ is the state at time t. We observe on-hand inventory level $I_t := \mathbf{s}_t(0)$ and the sales given by $y_t := \min(d_t, I_t)$, where $d_t \sim F$. Under the base-stock policy, the sales y_t also happen to be the order placed in the next time step. The new state at time $t+1$ is given by

$$\mathbf{s}_{t+1} = (s_t(0) - y_t + s_t(1), s_t(2), \ldots, s_t(L), y_t).$$

Let $n_T(\mathbf{s}_1) := \sum_{t=1}^{T} y_t$ denote the sum of sales from time 1 to T and $m_T(\mathbf{s}_1) := \sum_{t=1}^{T} I_t$ denote the sum of on-hand inventory levels.

Appendix C.1. Bounding Cumulative Observed Sales

We bound the difference between the total sales in time T starting from two different states $\mathbf{s}, \mathbf{s}'$ when the states satisfy the following property given.

Definition C.1. Define states $\mathbf{s} := (s(0), s(1), \ldots, s(L)), \mathbf{s}' := (s'(0), s'(1), \ldots, s'(L))$. We say that $s' \succeq s$ if $\mathbf{s}' = (s(0) + \delta_0, s(1) + \delta_1, \ldots, s(L) + \delta_L)$, where $\delta_0 + \delta_1 + \ldots + \delta_L = 0$ and there exists some $0 \le k \le L-1$ such that $\delta_i \ge 0$ for all $i \in \{0, 1, \ldots, k\}$ and $\delta_i \le 0$ for all $i \in \{k+1, k+2, \ldots, L\}$.

We first provide a simple bound on $n_T(\mathbf{s}_1') - n_T(\mathbf{s}_1)$ when $\mathbf{s}_1' \succeq \mathbf{s}_1$ and $T \le L+1$, which will be useful in our proof for larger T.

Lemma C.1. *Given two states $\mathbf{s}_1, \mathbf{s}_1'$ with $\mathbf{s}_1' \succeq \mathbf{s}_1$ and $\mathbf{s}_1, \mathbf{s}_1' \in \mathcal{S}^x$, define $Y_t := \sum_{i=1}^{t} y_t$, $Y_t' := \sum_{i=1}^{t} y_t'$ to be the total observed sales up to time t in process $\mathcal{M}(x, \mathbf{s}_1)$ and $\mathcal{M}(x, \mathbf{s}_1')$, respectively. Then, for $t = 1, 2, \ldots, L+1$, we have that*

$$Y_t' - Y_t \le \max_{0 \le k \le t-1} (\delta_0 + \ldots + \delta_k),$$

where $\delta_i = s'(i) - s(i)$ as defined in Definition C.1.

Proof. We couple the demand realizations in the two processes $\mathcal{M}(x, \mathbf{s}_1)$ and $\mathcal{M}(x, \mathbf{s}_1')$ so that the demands at time t are the same for both of the processes, denoted by d_t. We prove the lemma statement by induction on t. For $t = 1$,

$$y_1' - y_1 = \min(s(0) + \delta_0, d_1) - \min(s(0), d_1) \le \delta_0$$

because $\mathbf{s}_1' \succeq \mathbf{s}_1$ implies that $s(0) + \delta_0 \ge s(0)$. Assume for any time up to $t-1$, the hypothesis holds. Then, consider time $t \le L+1$, and observe that on-hand inventory is

$$I_t' = s_t'(0) = (s(0) + \delta_0 + s(1) + \delta_1 + \ldots + s(t-1) + \delta_{t-1}) - (y_1' + \ldots + y_{t-1}')$$

and

$$I_t = s_t(0) = (s(0) + s(1) + \ldots + s(t-1)) - (y_1 + \ldots + y_{t-1}),$$

so subtracting, we get

$$I_t' - I_t + Y_{t-1}' - Y_{t-1} = \delta_0 + \delta_1 + \ldots + \delta_{t-1}. \tag{C.1}$$

Now, we write

$$Y_t' - Y_t = y_t' - y_t + Y_{t-1}' - Y_{t-1} = \min(I_t', d_t) - \min(I_t, d_t) + Y_{t-1}' - Y_{t-1}.$$

There are four cases to consider.

1. $d_t \le I_t', d_t \le I_t$. In this case, $Y_t' - Y_t = d_t - d_t + Y_{t-1}' - Y_{t-1} = Y_{t-1}' - Y_{t-1} \le \max_{0 \le k \le t-2}(\delta_0 + \ldots + \delta_k) \le \max_{0 \le k \le t-1}(\delta_0 + \ldots + \delta_k)$ by the induction hypothesis.
2. $d_t \ge I_t', d_t \ge I_t$. In this case, $Y_t' - Y_t = I_t' - I_t + Y_{t-1}' - Y_{t-1} = \delta_0 + \ldots + \delta_{t-1} \le \max_{0 \le k \le t-1}(\delta_0 + \ldots + \delta_k)$ by (C.1).
3. $I_t \le d_t \le I_t'$. In this case, $Y_t' - Y_t = d_t - I_t + Y_{t-1}' - Y_{t-1} = d_t - I_t' + I_t' - I_t + Y_{t-1}' - Y_{t-1} = d_t - I_t' + \delta_0 + \ldots + \delta_{t-1} \le \delta_0 + \ldots + \delta_{t-1} \le \max_{0 \le k \le t-1}(\delta_0 + \ldots + \delta_k)$ by (C.1).
4. $I_t' \le d_t \le I_t$. In this case, $Y_t' - Y_t = I_t' - d_t + Y_{t-1}' - Y_{t-1} \le Y_{t-1}' - Y_{t-1} \le \max_{0 \le k \le t-2}(\delta_0 + \ldots + \delta_k) \le \max_{0 \le k \le t-1}(\delta_0 + \ldots + \delta_k)$ by the induction hypothesis.

Therefore, we have proven that under the induction hypothesis,

$$Y_t' - Y_t \le \max_{0 \le k \le t-1} (\delta_0 + \ldots + \delta_k),$$

and the desired result for all $t \in \{1, 2, \ldots, L+1\}$ follows by induction. □

Lemma C.2. *Consider the MRPs on following base-stock policy with level x starting in states $\mathbf{s}_1, \mathbf{s}_1' \in \mathcal{S}^x$ with $\mathbf{s}_1' \succeq \mathbf{s}_1$. Let $I_t = \mathbf{s}_t(0), I_t' = \mathbf{s}_t'(0)$ be the on-hand inventory levels in the two processes at time t. Then, if $I_t' - I_t \ge 0$ for all $t \in \{1, 2, \ldots, L+1\}$, then it holds that $n_T(\mathbf{s}_{L+1}') = n_T(\mathbf{s}_{L+1})$ for any T.*

Proof. If we have $I_t' - I_t \ge 0$, then the respective sales at time t satisfy $y_t' \ge y_t$ as well. Therefore, each entry of state $\mathbf{s}_{L+1}' = (I_{L+1}', y_1', y_2', \ldots, y_L')$ is at least the respective entry of state $\mathbf{s}_{L+1} = (I_{L+1}, y_1, y_2, \ldots, y_L)$. Because the total sum of the entries in each state is equal to x, we conclude that $\mathbf{s}_{L+1}' = \mathbf{s}_{L+1}$ and hence, $n_T(\mathbf{s}_{L+1}') = n_T(\mathbf{s}_{L+1})$ for any T. □

Lemma C.2 shows that if we ever observe a t with $\mathbf{s}_t' \succeq \mathbf{s}_t$, $\mathbf{s}_1, \mathbf{s}_1' \in \mathcal{S}^x$ and the next L consecutive on-hand inventory levels in the process starting from state $\mathbf{s}_t'$ are at least as high as when starting from state $\mathbf{s}_t$, then the two processes will reach an identical state at time $t+L$; hence, all future observed sales will be the same. Utilizing this property, for states $\mathbf{s}_1' \succeq \mathbf{s}_1$ we can define the following sequence of times at which the two processes get synchronized.

Definition C.2. Given starting states $\mathbf{s}_1, \mathbf{s}_1'$ of two coupled processes $\mathcal{M}(x, \mathbf{s}_1), \mathcal{M}(x, \mathbf{s}_1')$, with $\mathbf{s}_1' \succeq \mathbf{s}_1$, define a sequence of times

$$1 = \sigma_0 < \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \ldots \le \Gamma$$

as follows. For $i \ge 1$, τ_i is the first time after $t = \sigma_{i-1}$ at which $I_{\tau_i}' < I_{\tau_i}$, σ_i is the first time after $t = \tau_i$ at which $I_{\sigma_i}' > I_{\sigma_i}$, and Γ is the first time at which $\mathbf{s}_\Gamma' = \mathbf{s}_\Gamma$. By Lemma C.2, $\tau_i - \sigma_{i-1} \le L$ and $\sigma_i - \tau_i \le L$ (whenever τ_i, σ_i exist).

Lemma C.3. *Given starting states $\mathbf{s}_1' \succeq \mathbf{s}_1$ and the sequence defined (Definition C.2), $\mathbf{s}_{\sigma_i}' \succeq \mathbf{s}_{\sigma_i}$ and $\mathbf{s}_{\tau_i}' \preceq \mathbf{s}_{\tau_i}$ for all i, where $\tau_i, \sigma_i \le \Gamma$.*

Proof. We have $\mathbf{s}_{\sigma_0}' \succeq \mathbf{s}_{\sigma_0}$ for the starting state at time $t = 1 = \sigma_0$. If time $t = \tau_1 \le \Gamma$ exists, then $\tau_1 - \sigma_0 \le L$ by Lemma C.2. Furthermore, we can show that $\mathbf{s}_{\tau_1}' \preceq \mathbf{s}_{\tau_1}$. To see this,

$$\mathbf{s}_{\tau_1}' = (I_{\tau_1}', s_{\sigma_0}'(\tau_1 + 1 - \sigma_0), \ldots, s_{\sigma_0}'(L), y_{\sigma_0}', \ldots, y_{\tau_1 - 1}')$$

and

$$\mathbf{s}_{\tau_1} = (I_{\tau_1}, s_{\sigma_0}(\tau_1 + 1 - \sigma_0), \ldots, s_{\sigma_0}(L), y_{\sigma_0}, \ldots, y_{\tau_1 - 1}).$$

By definition of τ_1, for times $t \in \{\sigma_0, \sigma_0+1, \ldots, \tau_1-1\}$, we have $I'_t \geq I_t$, and hence, $y'_t \geq y_t$. We also know that $I'_{\tau_1} < I_{\tau_1}$. It suffices to show that $s'_{\sigma_0}(i) \leq s_{\sigma_0}(i)$ for all $i \in \{\tau_1+1-\sigma_0, \ldots, L\}$.

Recall $I'_{\tau_1} = I'_{\tau_1-1} - y'_{\tau_1-1} + s'_{\tau_1-1}(1) = I'_{\tau_1-1} - y'_{\tau_1-1} + s'_{\sigma_0}(\tau_1 - \sigma_0)$ and similarly, $I_{\tau_1} = I_{\tau_1-1} - y_{\tau_1-1} + s_{\sigma_0}(\tau_1-\sigma_0)$ so that

$$\begin{aligned} I'_{\tau_1-1} - y'_{\tau_1-1} + s'_{\sigma_0}(\tau_1-\sigma_0) &< I_{\tau_1-1} - y_{\tau_1-1} + s_{\sigma_0}(\tau_1-\sigma_0) \\ &\leq I'_{\tau_1-1} - y'_{\tau_1-1} + s_{\sigma_0}(\tau_1-\sigma_0), \end{aligned}$$

where the last inequality holds because $I'_{\tau_1-1} \geq I_{\tau_1-1}$ implies that (for any given demand d_{τ_1-1})

$$\begin{aligned} &I'_{\tau_1-1} - y'_{\tau_1-1} = I'_{\tau_1-1} - \min(I'_{\tau_1-1}, d_{\tau_1-1}) = (I'_{\tau_1-1} - d_{\tau_1-1})^+ \\ &\geq (I_{\tau_1-1} - d_{\tau_1-1})^+ = I_{\tau_1-1} - \min(I_{\tau_1-1}, d_{\tau_1-1}) = I_{\tau_1-1} - y_{\tau_1-1}. \end{aligned}$$

Hence, $s'_{\sigma_0}(\tau_1-\sigma_0) \leq s_{\sigma_0}(\tau_1-\sigma_0)$, and because $\mathbf{s}'_1 \succeq \mathbf{s}_1$, $s'_{\sigma_0}(i) \leq s_{\sigma_0}(i)$ holds for all $i \in \{\tau_1+1-\sigma_0, \ldots, L\}$. So, combined with the observation that $y'_t \geq y_t, t \in \{\sigma_0, \ldots, \tau_1-1\}$, we have shown that $\mathbf{s}'_{\tau_1} \preceq \mathbf{s}_{\tau_1}$.

We can inductively apply the argument for each successive σ_i, τ_i, so that $\mathbf{s}'_{\sigma_i} \succeq \mathbf{s}_{\sigma_i}$ and $\mathbf{s}'_{\tau_i} \preceq \mathbf{s}_{\tau_i}$ for all i. □

Finally, we are ready to bound the difference in total observed sales in time T starting from two states $\mathbf{s}'_1 \succeq \mathbf{s}_1$, with $\mathbf{s}_1, \mathbf{s}'_1 \in \mathcal{S}^x$, under base-stock policy with level x.

Lemma C.4. *Let $\mathbf{s}'_1, \mathbf{s}_1 \in \mathcal{S}^x$, and $\mathbf{s}'_1 \succeq \mathbf{s}_1$. Then,*

$$|n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1)| \leq 3x.$$

Proof. Let sequence $\sigma_0, \tau_1, \sigma_1, \ldots$ be the sequence of time indices as in Definition C.2. First, we show that

$$n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) \geq -2x.$$

Let us assume that in our sequence of times, the last σ is σ_M. Then, note that

$$\begin{aligned} n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) &= \sum_{t=1}^{T} y'_t - \sum_{t=1}^{T} y_t \\ &= \sum_{i=0}^{M-1} \left(\sum_{j=\sigma_i}^{\sigma_{i+1}-1} (y'_j - y_j) \right) + \sum_{t=\sigma_M}^{T} (y'_t - y_t). \end{aligned}$$

We will show that $\sum_{j=\sigma_i}^{\sigma_{i+1}-1} (y'_j - y_j) \geq 0$ for any $i = 0, 1, \ldots, M-1$. Consider the process starting from states $\mathbf{s}'_{\sigma_i}, \mathbf{s}_{\sigma_i}$, where $\mathbf{s}'_{\sigma_i} \succeq \mathbf{s}_{\sigma_i}$ by Lemma C.3.

By (C.1) in the proof of Lemma C.1,

$$\begin{aligned} &(y'_{\sigma_i} + \ldots + y'_{\tau_{i+1}-1}) - (y_{\sigma_i} + \ldots + y_{\tau_{i+1}-1}) \\ &= [(s'_{\sigma_i}(0) - s_{\sigma_i}(0)) + \ldots + (s'_{\sigma_i}(\tau_{i+1}-\sigma_i) - s_{\sigma_i}(\tau_{i+1}-\sigma_i))] \\ &\quad -(I'_{\tau_{i+1}} - I_{\tau_{i+1}}). \end{aligned} \tag{C.2}$$

Now, consider the (coupled) processes starting from states $\mathbf{s}'_{\tau_{i+1}}, \mathbf{s}_{\tau_{i+1}}$, where $\mathbf{s}'_{\tau_{i+1}} \preceq \mathbf{s}_{\tau_{i+1}}$. Recall that

$$\mathbf{s}'_{\tau_{i+1}} = (I'_{\tau_{i+1}}, s'_{\sigma_i}(\tau_{i+1}+1-\sigma_i), \ldots, s'_{\sigma_i}(L), y'_{\sigma_i}, \ldots, y'_{\tau_{i+1}-1})$$

and

$$\mathbf{s}_{\tau_{i+1}} = (I_{\tau_{i+1}}, s_{\sigma_i}(\tau_{i+1}+1-\sigma_i), \ldots, s_{\sigma_i}(L), y_{\sigma_i}, \ldots, y_{\tau_{i+1}-1}),$$

and as proved in Lemma C.3, $I'_{\tau_{i+1}} < I_{\tau_{i+1}}$, $s_{\sigma'_i}(i) \leq s_{\sigma_i}(i)$ for all $i \in \{\tau_{i+1}+1-\sigma_i, \ldots, L\}$ and $y'_t \geq y_t$ for all $t \in \{\sigma_i, \ldots, \tau_{i+1}-1\}$. So, $s_{\tau_{i+1}} \succeq s'_{\tau_{i+1}}$, and by Lemma C.1, we have that

$$\begin{aligned} &(y_{\tau_{i+1}} + \ldots + y_{\sigma_{i+1}-1}) - (y'_{\tau_{i+1}} + \ldots + y'_{\sigma_{i+1}-1}) \\ &\leq (I_{\tau_{i+1}} - I'_{\tau_{i+1}}) + [(s_{\sigma_i}(\tau_{i+1}+1-\sigma_i) \\ &-s'_{\sigma_i}(\tau_{i+1}+1-\sigma_i)) + \ldots + (s_{\sigma_i}(L) - s'_{\sigma_i}(L))] \\ &= (I_{\tau_{i+1}} - I'_{\tau_{i+1}}) + [(s'_{\sigma_i}(0) - s_{\sigma_i}(0)) + \ldots + (s'_{\sigma_i}(\tau_{i+1}-\sigma_i) - s_{\sigma_i}(\tau_{i+1}-\sigma_i))], \end{aligned} \tag{C.3}$$

where the last equality follows from the fact that because $\mathbf{s}_{\sigma_i}, \mathbf{s}'_{\sigma_i} \in \mathcal{S}^x$, the sum of all the entries in a state is always the same as x.

Combining the two results in (C.2) and (C.3), we have that for any $i = 0, 1, \ldots, M-1$,

$$\sum_{j=\sigma_i}^{\sigma_{i+1}-1} (y'_j - y_j) = (y'_{\sigma_i} + \ldots + y'_{\sigma_{i+1}-1}) - (y_{\sigma_i} + \ldots + y_{\sigma_{i+1}-1}) \geq 0.$$

Therefore, we can conclude that

$$n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) \geq \sum_{t=\sigma_M}^{T} (y'_t - y_t) = \sum_{t=\sigma_M}^{\hat{\Gamma}} (y'_t - y_t),$$

where $\hat{\Gamma} := \min(\Gamma, T)$. By our construction of the σ, τ sequence, $\hat{\Gamma} - \sigma_M + 1 \leq 2(L+1)$. Note that over any $L+1$ consecutive time steps, the total observed sales difference in those $L+1$ time steps can be at most x for any two starting states (this is because the total inventory position is always x). So, $n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) \geq \sum_{t=\tau_M}^{\hat{\Gamma}} (y'_t - y_t) \geq -2x$.

To complete the proof, we show in a similar way that $n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) \leq 3x$. Let us assume that in our sequence of times, the last τ is τ_K. Then, note that

$$\begin{aligned} n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) &= \sum_{t=1}^{T} y'_t - \sum_{t=1}^{T} y_t \\ &= \sum_{t=1}^{\tau_1-1} (y'_t - y_t) + \sum_{i=1}^{K-1} \left(\sum_{j=\tau_i}^{\tau_{i+1}-1} (y'_j - y_j) \right) + \sum_{t=\tau_K}^{T} (y'_t - y_t). \end{aligned}$$

For any $i = 1, 2, \ldots, K-1$, consider the process starting from states $\mathbf{s}'_{\tau_i}, \mathbf{s}_{\tau_i}$, where $\mathbf{s}'_{\tau_i} \preceq \mathbf{s}_{\tau_i}$ by the previous lemma. By an identical argument, we can show that $\sum_{j=\tau_i}^{\tau_{i+1}-1} (y'_j - y_j) \leq 0$ and $\sum_{t=\tau_K}^{T} (y'_t - y_t) = \sum_{t=\tau_K}^{\hat{\Gamma}} (y'_t - y_t) \leq 2x$. Noting that because there are at most $L+1$ time steps in $\sum_{t=1}^{\tau_1-1} (y'_t - y_t)$, it is bounded by x. Thus, we have shown that $n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1) \leq 3x$. Combining with the lower bound of $-2x$, we have

$$|n^x_T(\mathbf{s}'_1) - n^x_T(\mathbf{s}_1)| \leq 3x. \quad \square$$

Appendix C.2. Bounding the Cumulative On-Hand Inventory Level

Lemma C.5. *Let $\mathbf{s}', \mathbf{s} \in \mathcal{S}^x$, and $\mathbf{s}' \succeq \mathbf{s}$. Then,*

$$|m^x_T(\mathbf{s}) - m^x_T(\mathbf{s}')| \leq 6Lx.$$

Proof. Recall from before that y_t, I_t are the observed sales and the on-hand inventory level at the beginning of time $t \geq 1$, respectively. Under the base-stock level x policy, the order placed at time t is precisely y_t (assume without loss of generality $y_1 = 0$ because we start at a state $\mathbf{s}$ or $\mathbf{s}'$ with total inventory position x). Also, given starting state

$\mathbf{s} = (s(0), s(1), \ldots, s(L))$, we denote $y_0 := s(L), y_{-1} := s(L-1), \ldots, y_{1-L} := s(1)$. Because under the base-stock policy, the new order is equal to the sales, the on-hand inventory level transitions as follows:

$$I_{t+1} = I_t - y_t + y_{t-L}.$$

Therefore, we can write the on-hand inventory level at any time $k \geq 1$ as

$$I_k = I_1 + \sum_{j=1}^{k-1}(y_{j-L} - y_j),$$

and hence, the total sum of all on-hand inventory levels up to time T is

$$\begin{aligned}\sum_{k=1}^{T} I_k &= \sum_{k=1}^{T}\left(I_1 + \sum_{j=1}^{k-1}(y_{j-L} - y_j)\right)\\ &= TI_1 + \sum_{k=2}^{T}\sum_{j=1}^{k-1}(y_{j-L} - y_j)\\ &= TI_1 + \sum_{i=1}^{T-1}(T-i)(y_{i-L} - y_i)\\ &= TI_1 + \sum_{i=1}^{T-1}(T-i)y_{i-L} - \sum_{i=1}^{T-1}(T-i)y_i.\end{aligned}$$

Now, if we break up the summations on the right-hand side and reindex,

$$\begin{aligned}\sum_{i=1}^{T-1}(T-i)y_{i-L} &= \sum_{i=1}^{L}(T-i)y_{i-L} + \sum_{i=L+1}^{T-1}(T-i)y_{i-L}\\ &= \sum_{i=1}^{L}(T-i)y_{i-L} + \sum_{i=1}^{T-L-1}(T-L-i)y_i\end{aligned}$$

and

$$\sum_{i=1}^{T-1}(T-i)y_i = \sum_{i=1}^{T-L-1}(T-i)y_i + \sum_{i=T-L}^{T-1}(T-i)y_i,$$

so that

$$\begin{aligned}\sum_{k=1}^{T} I_k &= TI_1 + \sum_{i=1}^{T-1}(T-i)y_{i-L} - \sum_{i=1}^{T-1}(T-i)y_i\\ &= TI_1 + \sum_{i=1}^{L}(T-i)y_{i-L} - \left(\sum_{i=1}^{T-L-1} Ly_i + \sum_{i=T-L}^{T-1}(T-i)y_i\right).\end{aligned}$$

Define

$$\epsilon := \sum_{i=1}^{T-1} Ly_i - \left(\sum_{i=1}^{T-L-1} Ly_i + \sum_{i=T-L}^{T-1}(T-i)y_i\right) \geq 0,$$

and observe that

$$\epsilon = \sum_{i=1}^{L-1} iy_{T-L+i} \leq L\sum_{i=1}^{L-1} y_{T-L+i} \leq Lx \tag{C.4}$$

because in $L-1$ consecutive time steps, the total sales on following the base-stock level x policy cannot exceed x.

We can write

$$\begin{aligned}\sum_{k=1}^{T} I_k &= TI_1 + \sum_{i=1}^{L}(T-i)y_{i-L} - \sum_{i=1}^{T-1} Ly_i + \epsilon\\ &= \sum_{i=0}^{L}(T-i)s(i) - L\sum_{i=1}^{T-1} y_i + \epsilon\end{aligned}$$

by substituting the defined values of $y_0, \ldots, y_{1-L}$.

Now, let $I'_t, y'_t, s'(i), \epsilon'$ be the respective values if the starting state is $\mathbf{s}'$ instead of $\mathbf{s}$, with $\mathbf{s}' \geq \mathbf{s}$. Then, an expression similar to that given can be derived for $\sum_{k=1}^{T} I'_k$. Therefore, the difference $|m_x^T(\mathbf{s}') - m_x^T(\mathbf{s})|$ can be bounded as

$$\begin{aligned}|m_x^T(\mathbf{s}') - m_x^T(\mathbf{s})| &= \left|\sum_{k=1}^{T} I'_k - \sum_{k=1}^{T} I_k\right|\\ &\leq \left|\sum_{i=0}^{L}(T-i)s'(i) - \sum_{i=0}^{L}(T-i)s(i)\right|\\ &\quad + L\left|\sum_{i=1}^{T-1}(y'_i - y_i)\right| + |\epsilon' - \epsilon|\\ &\leq (Tx - (T-L)x) + L(3x) + 2(Lx)\\ &= 6Lx,\end{aligned}$$

where we bounded $|\sum_{i=0}^{L}(T-i)s'(i) - \sum_{i=0}^{L}(T-i)s(i)|$ by the largest possible value that occurs when $\mathbf{s}' = (x, 0, \ldots, 0)$ and $\mathbf{s} = (0, \ldots, 0, x)$, used (C.4) to bound $|\epsilon| \leq LX$, and used Lemma C.4 to bound $\sum_{i=1}^{T-1}(y'_i - y_i)$. □

Endnote

[1] Here, for simplicity of notation, we assume deterministic policies. The results in this section can be easily extended to randomized policies.

References

Agarwal A, Foster DP, Hsu DJ, Kakade SM, Rakhlin A (2011) Stochastic convex optimization with bandit feedback. Taylor JS, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems (NIPS 2011)*, Granada, Spain, 1035–1043.

Agrawal S, Jia R (2017) Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R, eds. *Adv. Neural Inform. Processing Systems 30 (NIPS 2017, Long Beach, CA)*, 1184–1194.

Bartlett PL, Tewari A (2009) REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. Bilmes JA, Ng AY, eds. *Proc. Twenty-Fifth Conf. Uncertainty Artificial Intelligence*, Montreal, QC, Canada (AUAI Press, Arlington, VA), 35–42.

Bartók G, Foster DP, Pál D, Rakhlin A, Szepesvári C (2014) Partial monitoring—classification, regret bounds, and algorithms. *Math. Oper. Res.* 39(4):967–997.

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Sci.* 59(6): 1407–1424.

Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.

Bijvank M, Vis IF (2011) Lost-sales inventory theory: A review. *Eur. J. Oper. Res.* 215(1):1–13.

Huh WT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Math. Oper. Res.* 34(1):103–123.

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009a) An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Math. Oper. Res.* 34(2):397–416.

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009b) Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Sci.* 55(3):404–420.

Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *J. Machine Learn. Res.* 11(Apr):1563–1600.

Janakiraman G, Roundy RO (2004) Lost-sales problems with stochastic lead times: Convexity results for base-stock policies. *Oper. Res.* 52(5):795–803.

Lee HL, Cohen MA (1983) A note on the convexity of performance measures of m/m/c queueing systems. *J. Appl. Probab.* 20(4):920–923.

Lugosi G, Markakis MG, Neu G (2017) On the hardness of inventory management with censored demand data. Preprint, submitted October 16, https://arxiv.org/abs/1710.05739.

Puterman ML (2014) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, Hoboken, NJ).

Shanthikumar JG, Yao DD (1987) Optimal server allocation in a system of multi-server stations. *Management Sci.* 33(9):1173–1180.

Tewari A, Bartlett PL (2008) Optimistic linear programming gives logarithmic regret for irreducible MDPs. Platt JC, Koller D, Singer Y, Roweis ST, eds. *Proc. Twenty-First Annual Conf. Adv. Neural Inform. Processing Systems (NIPS 2007*, Vancouver, British Columbia, Canada) (Curran Associates, Inc.), 1505–1512.

Weber RR (1980) Note—On the marginal benefit of adding servers to g/gi/m queues. *Management Sci.* 26(9):946–951.

Zhang H, Chao X, Shi C (2020) Closing the gap: A learning algorithm for the lost-sales inventory system with lead times. *Management Sci.* 66(5):1962–1980.

Zipkin P (2000) *Foundations of Inventory Management* (McGraw-Hill, Boston).

Zipkin P (2008) Old and new methods for lost-sales inventory systems. *Oper. Res.* 56(5):1256–1263.

Shipra Agrawal is an associate professor in the Department of Industrial Engineering and Operations Research at Columbia University. She is also affiliated with the Department of Computer Science and the Data Science Institute at Columbia University. Her research spans several areas of optimization, machine learning, and decision making, including online learning and decision making, multiarmed bandits, and reinforcement learning.

Randy Jia is a research scientist at Amazon in New York, NY. He recently earned his PhD in operations research from Columbia University. He is interested in machine learning and optimization, in particular multiarmed bandits and reinforcement learning.