Closing the Gap: A Learning Algorithm for Lost-Sales Inventory Systems with Lead Times

Huanan Zhang
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering,
Pennsylvania State University, University Park, PA 16802, huz157@psu.edu

Xiuli Chao, Cong Shi
Industrial and Operations Engineering, University of Michigan, MI 48105, {xchao, shicong}@umich.edu

We consider a periodic-review single-product inventory system with lost-sales and positive lead times under censored demand. In contrast to the classical inventory literature, we assume the firm does not know the demand distribution *a priori*, and makes adaptive inventory ordering decision in each period based only on the past sales (censored demand) data. The standard performance measure is regret, which is the cost difference between a feasible learning algorithm and the clairvoyant (full-information) benchmark. When the benchmark is chosen to be the (full-information) optimal base-stock policy, Huh et al. [*Mathematics of Operations Research* 34(2): 397-416 (2009)] developed a nonparametric learning algorithm with a cubic-root convergence rate on regret. An important open question is whether there exists a nonparametric learning algorithm whose regret rate matches the theoretical lower bound of any learning algorithms. In this work, we provide an affirmative answer to the above question. More precisely, we propose a new nonparametric algorithm termed *the simulated cycle-update policy*, and establish a square-root convergence rate on regret, which is proven to be the lower bound of any learning algorithms. Our algorithm uses a random cycle-updating rule based on an *auxiliary simulated system* running in parallel, and also involves two new concepts, namely, *the withheld on-hand inventory* and *the double-phase cycle gradient estimation.* The techniques developed are effective for learning a stochastic system with complex systems dynamics and lasting impact of decisions.

Key words: inventory, lost-sales, lead time, base-stock policy, censored demand, nonparametric, learning algorithms, regret analysis

Submitted February 2017; Revised March 2018, September 2018; Accepted December 2018.

1. Introduction

The periodic-review inventory control problem with lost-sales and positive lead times is one of the most fundamental yet notoriously difficult problems in the theory of inventory management (see Zipkin (2000)). The model assumes that unmet demand at the end of each period is *lost*, rather than being backlogged and carried over to the next period. For example, in many retail applications demand can be met by competing suppliers, making lost-sales a more appropriate modeling assumption (cf. Bijvank and Vis (2011)). There is a constant delivery lead time measured by the delay between placing an order and receiving it, which leads to an enlarged state-space in which the pipeline orders need to be tracked (cf. Zipkin (2008b)). In this paper, contrary to the classical inventory setting, we assume that the firm does not know the demand distribution *a priori* but can only collect past sales data over time. Because the sales in a period are the minimum of the actual demand and the on-hand inventory level, the demand information is *censored* (cf. Huh et al. (2009a)). The firm wishes to minimize the long-run average holding and lost-sales penalty cost per period.

Even with complete information about the demand distribution, it is well-known that the optimal policy does not possess a simple form (see Karlin and Scarf (1958), Morton (1969), Janakiraman and Roundy (2004), Janakiraman et al. (2007)). To analyze the structure of the optimal policies, Zipkin (2008b) used a partial sum of inventory to represent the state and showed that the minimum cost function is $L^\natural$-convex, and as a result, the optimal order quantities exhibit monotonicity and bounded sensitivity (more sensitive to newer orders). Although analyzing the dynamic program with large state space yields such nice structural properties, the computation of optimal policies remains intractable due to the well-known *curse of dimensionality.* As a result, a considerable amount of efforts has been devoted to designing various effective heuristic policies (Reiman (2004), Levi et al. (2008), Zipkin (2008a), Lu et al. (2015), Goldberg et al. (2016), Xin and Goldberg (2016)). In particular, Huh et al. (2009b) showed that the best base-stock policy is an effective heuristic. Levi et al. (2008) proposed a dual-balancing policy for this problem so that the expected cost of their policy is always within two times the expected optimal cost, and Chen et al. (2014) applied the $L^\natural$-convexity results to devise a pseudo-polynomial time approximation scheme that solves this problem within an arbitrary prespecified additive error. More recently, Xin and Goldberg (2016) showed that the best constant-order policy converges to optimality exponentially fast as lead time grows large.

As we have witnessed the recent progress for this fundamental class of problems, the incomplete information counterpart problem (under censored demand) remains relatively under-explored. In many practical scenarios (e.g., furniture retailing), the firm does not know the underlying demand distribution *a priori* and is forced to make replenishment decisions based on historical sales data. However, the sales data, as we discussed earlier, are in fact censored demand information. The joint learning and optimization problem in the underlying lost-sales system is therefore practically relevant and theoretically challenging. The only paper (and the closest to ours) in the literature is Huh et al. (2009a) who studied the exact same model and proposed an online learning algorithm whose regret against the full-information *optimal base-stock policy* is $O(T^{2/3})$ over a T-period problem. The motivations and justifications for using the optimal base-stock policy as a valid benchmark for this incomplete information problem are two-fold. First, the class of base-stock policies is easily implemented and widely used (see e.g., Janakiraman and Roundy (2004)). Second, Huh et al. (2009b) showed that, with complete information, as the unit penalty cost increases, with other parameters unchanged, the ratio of the cost of the best base-stock policy to the optimal cost converges to one. Their numerical results suggest "*when the ratio between the lost-sales penalty and the holding cost is 100, the cost of the best base-stock policy typically is within 1.5% of the optimal cost*". In many applications, this ratio "typically exceeds 200" (see Huh et al. (2009a)). We also refer interested readers to Bijvank et al. (2014) for a robustness result on the asymptotic optimality of base-stock policy in lost-sales inventory systems. Therefore, the class of base-stock policies is expected to perform very well.

An important open question raised by Huh et al. (2009a) is that whether there exists a nonparametric learning algorithm whose regret matches the theoretical lower bound $\Omega(\sqrt{T})$.

1.1. Main Results and Contributions

This paper provides an affirmative answer to the open question left by Huh et al. (2009a). More specifically, for the periodic-review inventory control problem with lost-sales and a positive lead time $L \geq 1$ under censored demand information, we present a new nonparametric learning algorithm, termed the *simulated cycle-update algorithm* (SCU for short), and show that the expected regret, defined as the difference in cost between the SCU and the optimal base-stock policy, is on the order of $O(\sqrt{T})$ for a T-period problem, which matches the theoretical lower bound (see Theorem 1 and Proposition 1). Our numerical results also show that the SCU algorithm performs better than the learning algorithm proposed in Huh et al. (2009a).

The SCU algorithm belongs to the broad family of *online gradient decent* (OGD) type of algorithms developed for various other inventory systems (cf. Burnetas and Smith (2000), Huh and Rusmevichientong (2009), Shi et al. (2016), Zhang et al. (2018)). Most studies on lost-sales inventory systems, with the exception of Huh et al. (2009a), considered models with zero lead times, that are significantly easier to analyze. One major challenge is that with positive lead times, each order placed has a prolonged impact (for at least a lead time of L periods) on the state of the system as well as the cost. Conventional online learning algorithms in the literature cannot be readily adapted to such a stochastic system, due to this lasting impact on decision-making and the complex system dynamics.

To tackle the aforementioned challenge, at a high-level, we develop a random cycle-updating rule (on the base-stock levels) based on another simulated system running in parallel, so that the (prolonged) cost impact of revising a target base-stock level can be readily quantified and compared between two feasible policies. Next, we highlight the main novelties of our approach below.

(a) First, our SCU algorithm cyclically updates base-stock level in a subset of periods termed the *triggering periods*. More specifically, the triggering periods are sequentially determined whenever another parallel auxiliary simulated system (operating under a lower base-stock level) experiences no lost-sales for L consecutive periods, then it triggers the beginning of a new cycle. The intuition is as follows. Consider two systems operating under different base-stock levels, if both systems experience no lost-sales for L consecutive periods, then the difference in state between these two systems would be only in the on-hand inventory, as both systems would share the same pipeline inventories. As a result, we can effectively compare the costs of any two feasible policies within a cycle (between two consecutive triggering periods). It can also be shown that the cost of any feasible base-stock policy within each cycle is convex with respect to the base-stock level. Note that this needed convexity result does not hold for pre-determined fixed cycles, where the initial state at the start of each cycle remains unknown.

(b) The second idea is that we introduce a new concept termed *withheld on-hand inventory* in which we iteratively temporarily mark off some inventory units (according to a well-defined rule). The purpose of introducing this concept is to trigger ordering decisions that allow us effectively learn about demand. Note that we are not throwing these withheld inventory units away, but rather we pretend them to be nonexistent when ordering decisions are made. We use the withheld on-hand inventory to serve demand only when all the other on-hand inventory has been consumed. The rationale is as follows. By temporarily marking off these inventory units, we will order minimum extra inventory to maintain no less on-hand inventory than the simulated system, thereby allowing us to gather sufficient demand information to keep the simulated system running properly. If it happens that the on-hand inventory is less than the simulated system, then when our system experiences a lost-sales, we will be unable to determine if the simulated system also experiences a lost-sale or not. While having the withheld on-hand inventory is necessary to run the simulated system, we show that the additional average regret introduced by the withheld on-hand inventory is bounded by $O(\sqrt{T})$, so that it does not affect the overall regret bound.

(c) The third idea is that we use a double-phase approach to obtain a biased but good cost gradient estimator. The reason for introducing this new approach is that, with positive lead times, whenever we revise the base-stock level, it is not possible to immediately adjust the on-hand inventory level and therefore we may not have enough demand information to extract the cost gradient within a cycle. The cost gradient obtained by our double-phase approach is subject to estimation bias. However, we show that this estimation bias vanishes by establishing some convergence results for Markov chains with continuous state space (or Harris chains).

There are key differences between the present paper and Huh et al. (2009a). The first difference is the cycles constructed: The cycles in Huh et al. (2009a) are pre-determined and increasing in length (with cycle k containing $\lceil\sqrt{k}\rceil$ periods), while the cycles in the SCU algorithm have random lengths. The second difference is in the gradient estimation: The gradient estimate in SCU is based on data from the second phase of each cycle, while the gradient estimate in Huh et al. (2009a) only uses demand information from *one period* of each cycle (considering the fact that their cycle lengths are increasing). Result-wise, the main improvement is that the regret upper bound of the SCU algorithm matches the theoretical lower bound of any learning algorithms.

1.2. Outline and General Notation

The rest of this paper is organized as follows. In §2, we formally describe the periodic-review inventory systems with lost-sales and positive lead times under censored demand information. In §3, we introduce the simulated cycle-update (SCU) algorithm and offer a detailed discussion on the main ideas underlying its algorithmic design. In §4, we analyze the performance of SCU and discuss how to change SCU to achieve a better numerical performance when the demand is uncensored (see §4.4). In §5, we test the empirical performance of SCU against the algorithm proposed in Huh et al. (2009a) as well as the uncensored counterpart algorithm of SCU proposed in §4.4. Finally, we conclude the paper and point out some future research directions in §6.

For any real numbers x and y, we denote $x^+ = \max\{x, 0\}$. The indicator function $\mathbb{1}(A)$ takes value 1 if A is true and 0 otherwise. The projection function is defined as $\mathbf{P}_{[a,b]}(x) = \min[b, \max(x, a)]$ for any real numbers x, a, b. For any real-valued vector $\mathbf{x}$, we use $\sum \mathbf{x}$ to denote the sum of all its entries. For example, if $\mathbf{x}$ is an L-dimensional vector $\mathbf{x} = (x_1, \dots, x_L)$, then $\sum \mathbf{x} = \sum_{i=1}^{L} x_i$.

2. Model Description

Consider a periodic-review inventory system with lost-sales, positive ordering lead times and censored demand. The demands over periods $\{D_1, D_2, \dots, D_t, \dots\}$ are i.i.d. continuous random variables. Let t denote the period, $t = 1, 2, \dots$, and let D denote a generic one-period demand, which is non-negative with $\mathbb{E}[D] > 0$. The ordering lead time is a fixed integer $L \geq 1$. Contrary to the classical formulation, the firm has no access to the true demand distribution *a priori*. The firm can only observe the past censored demand data and adjust the ordering decisions on the fly.

For the lost-sales inventory system under consideration, any new order will stay in the pipeline for L periods before arrival. Hence, together with the on-hand inventory, we need to use an $(L+1)$-dimensional vector to keep track of the inventory information. For every period t, the starting inventory, or state of the system, is denoted by

$$\mathbf{x}_t = [q_{t-1}, \dots, q_{t-L+1}, I_t],$$

where I_t is the on-hand inventory at the beginning of period t, and q_k is the order placed in period k. Let $\mathbf{y}_t = [q_t, q_{t-1}, \ldots, q_{t-L+1}, I_t]$ denote the inventory after ordering in period t. Clearly, all the entries of $\mathbf{x}_t$ and $\mathbf{y}_t$ are non-negative. For simplicity, let $q_k = 0$ for all $k \le 0$.

For any feasible policy π, the sequence of events in each period t, $t = 1, 2, \ldots$, is as follows. (Note that all the states and decisions depend on π, but in general we shall make the dependency implicit for notational simplicity. However, whenever necessary, we use $\mathbf{x}_t^\pi$ and q_t^π to represent the state and ordering decision of policy π in period t.)

(i) At the beginning of each period t, the firm observes the starting inventory vector $\mathbf{x}_t = [q_{t-1}, \ldots, q_{t-L+1}, I_t]$, and makes a replenishment decision $q_t \ge 0$.

(ii) Then, the demand D_t is realized, and we denote its realization by d_t. The demand is satisfied to the maximum extent by on-hand inventory I_t. Since demand is censored, the firm only observes sales quantity $\min(d_t, I_t)$. Thus, if $d_t \ge I_t$, then the firm does not know the exact demand.

(iii) At the end of the period, each remaining on-hand inventory unit incurs a per-unit holding cost h, and each unsatisfied demand unit incurs a per-unit lost-sales penalty cost p. As a result, the cost in period t, denoted by C_t^π, is

$$C_t^\pi = h(I_t - d_t)^+ + p(d_t - I_t)^+.$$

Note that the lost-sales quantity and its penalty cost (as an opportunity cost) are unobservable to the firm due to demand censoring.

(iv) At last, the system proceeds to period $t+1$ with system state $\mathbf{x}_{t+1}$ given by

$$\mathbf{x}_{t+1} = \left[q_t, \ldots, q_{t-L+2}, I_{t+1} = q_{t-L+1} + (I_t - d_t)^+\right]. \tag{1}$$

The objective is to find an ordering policy, based on historical sales information, that minimizes the expected average cost of the lost-sales inventory system with positive lead times.

As seen from §1, even when the demand distribution is known, the computation of an optimal policy is intractable due to the curse of dimensionality. When the demand distribution is not known *a prori*, it becomes even harder if we use the optimal policy as the benchmark. Hence in this paper, we follow Huh et al. (2009a) to use the best base-stock policy as the benchmark. The class of base-stock policies is parametrized by a single parameter, the base-stock level $S \ge 0$. Under a base-stock policy with a base-stock level S, the ordering quantity in period t is $q_t = \left(S - I_t - \sum_{i=t-L+1}^{t-1} q_i\right)^+$. Note that $I_t + \sum_{i=t-L+1}^{t-1} q_i$ is the inventory position at the beginning of period t. Thus, essentially, the base-stock policy orders to raise the inventory position to S if the starting inventory position is less than S, and orders nothing otherwise. We refer to Huh et al. (2009b) and Huh et al. (2009a) for the asymptotic optimality and the effectiveness of base-stock policies.

In this paper, we will design an adaptive learning inventory policy that only uses the past sales data, and show that the expected average cost of the policy converges to that of the optimal base-stock policy at rate $O(1/\sqrt{T})$, which matches the theoretical lower bound.

3. Nonparametric Algorithm - Simulated Cycle-Update Policy (SCU)

We present a learning algorithm which we refer to as *simulated cycle-update policy* (SCU for short). Before introducing the SCU policy, we make the following assumption on the optimal (full information) base-stock level S^*.

Assumption 1. *There exist two known finite numbers $\underline{D}$ and $\bar{D}$ with $\underline{D} < \bar{D}$, such that*

(i) $\mathbb{P}(D \le \underline{D}) = c_1 > 0$,

(ii) there exists a constant $\delta > 0$, $\mathbb{P}\left(D \ge \frac{S^*+\delta}{L+1}\right) = c_2 > 0$, *and*

(iii) $S^* \in [(L+1)\cdot\underline{D},\ (L+1)\cdot\bar{D}]$.

Assumption 1(iii) gives an upper and a lower bound on the optimal base-stock level S^*, which is a predominant and standard assumption in the nonparametric learning literature in inventory management (see, e.g., Huh and Rusmevichientong (2009), Huh et al. (2009a), Shi et al. (2016), Zhang et al. (2018)). Assumption 1(i) means that there is a positive probability that the demand falls below $\underline{D}$, which is very mild and also used in Huh et al. (2009a). Assumption 1(ii) roughly states that, under the optimal base-stock policy, there is a likelihood of lost-sales, so the system does not *always* have sufficient inventory. Although this assumption does not appear in Huh et al. (2009a), it is very mild and satisfied for almost all practical systems with random demand (unless $h = 0$ or $p = \infty$). For convenience, in what follows we let $\underline{S} = (L+1) = (L+1)\cdot\underline{D}$ and $\bar{S} = (L+1)\cdot\bar{D}$.

For ease of notation in our regret analysis, we shall assume, without loss of generality, that $\delta = 1$. This is because, without affecting the system cost, we can change the unit of the demand and inventory levels, and correspondingly change the unit cost parameters, to obtain an equivalent system with $\delta = 1$. To see this, we define a new system with $D^{\mathbf{NEW}} = D/\delta$, $S^{\mathbf{NEW}*} = S^*/\delta$ and cost parameters $h^{\mathbf{NEW}} = h\cdot\delta$, $p^{\mathbf{NEW}*} = p\cdot\delta$, then $\delta^{\mathbf{NEW}} = 1$ for the new system. The cost of the system operating under any feasible policy is the same for the two systems, and so is the regret bound. Therefore, for the rest of the regret analysis, we will rewrite Assumption 1(ii) as

$$\mathbb{P}\left(D \ge \frac{S^*+1}{L+1}\right) = c_2 > 0.$$

3.1. Random Cycles, the Simulated System, and the Function G

One of the main challenges in designing our algorithm is that the total cost of the system cannot be readily written in a form that is amenable for online optimization. To overcome this, our first step is to divide the time periods into appropriately designed learning cycles, and then update the inventory target levels from cycle to cycle (instead of from period to period). That is, we use the (censored) demand information collected from one particular cycle to update the base-stock level for its subsequent cycle.

3.1.1. Random cycles based on the simulated system. As its name suggests, the SCU algorithm is designed based on a concurrently simulated inventory system. This system is run in the background and it implements a base-stock policy $\underline{S}$. For convenience, we shall refer to this simulated system as the simulated $\underline{S}$-system. In what follows, we shall define a sequence of cycles using the simulated $\underline{S}$-system; and the SCU algorithm updates the base-stock level at the beginning of each cycle using data collected from the SCU-system in the previous cycle.

Specifically, we define a "triggering event" as the event that simulated $\underline{S}$-system experiences no lost-sales for L consecutive periods. We call the period after a triggering event a triggering period. Let t_k denote the k-th triggering period, and for convenience, we let period 1 be the first triggering period. Mathematically, the triggering periods are defined by

$$t_1 = 1, \quad t_{k+1} = \min\left\{ n \;\middle|\; n \geq t_k + L,\ I_i^{\underline{S}} > D_i,\ \text{for all } n - L \leq i < n \right\},$$

where the $I_i^{\underline{S}}$ is the on-hand inventory level of the simulated $\underline{S}$-system in period i. Note that in this definition, once a triggering period is found, it resets the counter for the consecutive number of lost-sales periods to zero. Huh et al. (2009a) have shown that, the on-hand inventory of a lost-sales inventory system with positive lead times is non-decreasing in its base-stock levels. This implies that, if t is a triggering period, then it is also a triggering period for the inventory system operating under any base-stock policy $S \geq \underline{S}$, and therefore the pipeline inventory under any base-stock policy $S \geq \underline{S}$ is $(d_{t-1}, d_{t-2}, \ldots, d_{t-L})$.

The cycles are defined as follows: Let τ_k be the first period of cycle k, $k = 1, 2, \ldots$, then $\tau_1 = t_1 = 1$, and for $k > 1$, $\tau_k = t_{2(k-1)}$. That is, the first cycle starts in period 1, and starting from the second cycle, each cycle contains two phases, and each phase begins with a triggering period. Let τ_k' denote the first period of the second phase of cycle k, then $\tau_k' = t_{2k-1}$, $k = 2, 3, \ldots$, as depicted in Figure 1. Note that the cycle length is *a priori* random, it is *independent* of the learning algorithm. The double-phase cycle is designed to overcome the difficulty that we may not be able to evaluate the first phase's stochastic gradient of the function G (defined below) that will be used to guide the update of base-stock levels (see more detailed discussion in §3.3.4).

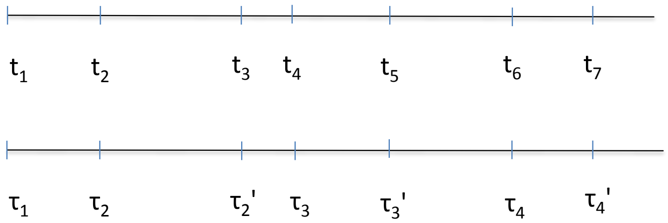

Figure 1 For $k \geq 1$, t_k is a triggering period, and τ_k and τ_k' are first periods of the two phases of cycle $k \geq 2$.

3.1.2. The function G. Next, we define an important function, $G(S,a,b)$, which denotes the total cost from period a to period b (both included), by using a base-stock level $S \geq \underline{S}$, and its starting state is specified as follows: If $a \geq L+1$ then we assume that the starting state in period a is $\left[d_{a-1}, d_{a-2}, \ldots, d_{a-L}, S - \sum_{i=a-L}^{a-1} d_i\right]$, otherwise the starting state is $[S, 0, \ldots, 0]$ in period a. Note that the function $G(S,a,b)$ also clearly depends on $(d_{a-1}, d_{a-2}, \ldots, d_{a-L})$, but we will make this dependency implicit for notational simplicity. We shall only consider the vector $(d_{a-1}, d_{a-2}, \ldots, d_{a-L})$ satisfying $\sum_{t=1}^{L} d_{a-t} \leq \underline{S}$. By Theorem 8 in Janakiraman and Roundy (2004), we know that $G(S,a,b)$ is convex in the base-stock level S, and is differentiable except for finite points. Let $\nabla G(S,a,b)$ denote the *partial* derivative of $G(S,a,b)$ with respect to S. The computation of $\nabla G(S,a,b)$ is discussed in §3.4.

3.2. The Simulated Cycle-Update (SCU) Policy

Before presenting the detailed algorithm, we first give a very high-level description of it, while deferring a detailed discussion of several new ideas involved in the algorithm to §3.3. The SCU algorithm proceeds in cycles, and recommends a new target inventory position at the start of each cycle. As discussed earlier in §3.1, the cycle is sequentially determined by what-we-call triggering periods in which another parallel auxiliary simulated $\underline{S}$-system experiences no lost-sales for L consecutive periods (since the last triggering period). Since $\underline{S}$-system is operated under a base-stock policy $\underline{S}$, any system operating under base-stock level $S \geq \underline{S}$ (including the optimal system) also experiences no lost-sales for L consecutive periods. Hence, the cycles are well-aligned between any two feasible systems. More importantly, whenever a trigger period occurs, the pipeline inventory becomes the same (involving only past L period demands) for any two feasible systems. This allows us to find the cycle cost gradient information about the function G defined §3.1 with respect to the base-stock level only, which is then used to update the target inventory position for the subsequent cycle.

However, there are two main sources of difficulties in evaluating the cycle cost gradient of G-system due to the transition from the existing target inventory position S_{k-1} to a new one S_k. On a high-level, when $S_k < S_{k-1}$ (i.e., pushing down the target inventory position), we cannot naively implement the base-stock policy S_k because it may cause the inability to simulate the $\underline{S}$-system (due to demand censoring). Hence, the algorithm needs to keep sufficient quantity to guarantee the simulation of the $\underline{S}$-system. To this end, we introduce an important new concept called the *withheld inventory* to keep track of the excess on-hand inventory caused by over-ordering above S_k. On the other hand, when $S_k \geq S_{k-1}$ (i.e., pushing up the target inventory position), we will need to introduce a two-phase design, i.e., each cycle contains two phases, each starting with a triggering period. This is because during the first phase, the gradient information of G-system may not be extracted (due to demand censoring), since the on-hand inventory level of the G-system could be higher than that of the SCU-system. However, it is viable to obtain such gradient information during the second phase. More detailed reasoning behind these two new concepts (i.e., the withheld inventory and the two-phase design) is given in §3.3.

With the necessary definitions in place, we present the detailed SCU algorithm. For convenience, we let $\tau_1' = \tau_1 = 1$. Let the step-size $\eta_k = \gamma/\sqrt{k}$ for all $k = 1, 2, \ldots$, for some positive constant γ. Note that since the cycle cost function $G(\cdot, a, b)$ (to be estimated) is convex, this choice of the stochastic gradient descent step is consistent with the literature on robust stochastic approximation (Nemirovski et al. (2009)) and online convex optimization (Hazan (2016)).

In each period t, the algorithm divides the total on-hand inventory I_t^{SCU} into two parts, namely, the withheld on-hand inventory denoted by $\hat{I}_t^{SCU}$, and the regular (or non-withheld) on-hand inventory denoted by $\tilde{I}_t^{SCU}$. The detailed evolution of the withheld inventory is given explicitly in the description of the algorithm. For every period t, we have $\hat{I}_t^{SCU} + \tilde{I}_t^{SCU} = \sum \mathbf{x}^{SCU}$.

Algorithm 1: Simulated Cycle-Update Algorithm (SCU)

Step 0 (Initialization):

- Start with an arbitrary target base-stock level $S_1 \in [\underline{S}, \bar{S}]$.
- Initialize the withheld on-hand inventory $\hat{I}_1^{SCU} = 0$.
- Set the initial inventory of both the SCU- and the simulated $\underline{S}$-systems to $\mathbf{x}_1^{SCU} = \mathbf{x}_1^{\underline{S}} = \mathbf{0}$.
- Set the counter for consecutive no lost-sales events for the simulated system $\psi = 0$. (Recall that cycles are defined using the simulated $\underline{S}$-system. In our SCU algorithm, we use ψ to record the number of consecutive no lost-sales periods in the simulated $\underline{S}$-system. When ψ reaches L, it signals a triggering period and ψ is reset to 0.)

Step 1: For the first cycle $k = 1$ starting with period $t = 1$, do the following.

1(a). Order q_t^{SCU} for the SCU system and $q_t^{\underline{S}}$ for the simulated $\underline{S}$ system as follows:

$$q_t^{SCU} = \left(S_k - \sum \mathbf{x}_t^{SCU} + \hat{I}_t^{SCU}\right)^+, \tag{2}$$

$$q_t^{\underline{S}} = \left(\underline{S} - \sum \mathbf{x}_t^{\underline{S}}\right)^+. \tag{3}$$

The ordering decision in the SCU-system is given as follows: It implements the modified base-stock policy S_k based on the regular inventory only (by temporarily ignoring the withheld inventory $\hat{I}_t^{SCU}$). More precisely, since the regular (or non-withheld) inventory position is $\sum \mathbf{x}_t^{SCU} - \hat{I}_t^{SCU}$, we order q_t^{SCU} of (2) to raise the regular inventory to S_k.

1(b). Observe the sales quantity $\min(d_t, I_t^{SCU})$, and update the withheld on-hand inventory by

$$\hat{I}_{t+1}^{SCU} := \left[\hat{I}_t^{SCU} - \left(\min(d_t, I_t^{SCU}) - \tilde{I}_t^{SCU}\right)^+\right]^+. \tag{4}$$

The demand fulfillment rule for the SCU-system is given as follows: It first uses the regular (or non-withheld) on-hand inventory to satisfy demand, and then uses the withheld on-hand inventory to satisfy demand (only after the regular on-hand inventory has been fully consumed). Thus, we update the withheld on-hand inventory following (4).

1(c). Update the states of both the SCU system and the simulated $\underline{S}$ system following the system dynamics (1), with the demand in period t for the simulated $\underline{S}$-system being replaced by $\min(d_t, I_t^{SCU})$.

1(d). If there is no lost-sales in the simulated $\underline{S}$ system, set the counter for consecutive no lost-sales events for the simulated system $\psi := \psi + 1$. Otherwise reset $\psi := 0$.

1(e). If $\psi = L$, then label period $t+1$ as a triggering period and reset $\psi = 0$. The sales data is used to compute $\nabla G(S_1, 1, t)$ following a well-defined subroutine presented in §3.4, and update the base-stock level S_2 for the second cycle as

$$S_2 = \mathbf{P}_{[\underline{S}, \bar{S}]}\left(S_1 - \eta_1 \nabla G(S_1, 1, t)\right).$$

Set $\tau_2 := t+1$, and update the withheld on-hand inventory by

$$\hat{I}_{\tau_2}^{SCU} := \left(\hat{I}_{\tau_2}^{SCU} - (S_2 - S_1)\right)^+,$$

and proceed to Step 2 with $k = 2$. On the other hand, if $\psi < L$, then repeat procedures 1(a) to 1(e) with $t := t+1$ if $t < T$, and stop otherwise.

Step 2: For cycles $k \geq 2$, each cycle contains two phases.

Phase 1: Start from period $t = \tau_k$.

2(a) Conduct procedures 1(a)–1(d) in Step 1.

2(b) If $\psi = L$, then set $\tau_k' = t+1$ and $\psi = 0$, and proceed to Phase 2. Otherwise, repeat 2(a) with $t := t+1$ if $t < T$, and stop otherwise.

Phase 2: Start from period $t = \tau'_k$.

2(a') Conduct procedures 1(a)–1(d) in Step 1.

2(b') If $\psi = L$, then set $\tau_{k+1} := t+1$. Update the target base-stock level for the next cycle as

$$S_{k+1} = \mathbf{P}_{[\underline{S},\bar{S}]}\left(S_k - 2\eta_k \nabla G\left(S_k, \tau'_k, t\right)\right),$$

Note that here we double the gradient of the second phase to estimate the gradient of the whole cycle. We then update the withheld on-hand inventory by

$$\hat{I}_t^{SCU} := \left(\hat{I}_t^{SCU} - (S_{k+1} - S_k)\right)^+.$$

Set $\psi := 0$, $k := k+1$, and repeat Step 2. If $\psi < L$, then repeat 2(a') with $t := t+1$ if $t < T$, and stop otherwise.

This concludes the description of the SCU algorithm.

3.3. Main Ideas of the SCU Algorithm

The SCU algorithm involves several main ideas, and we have discussed one of them, which is the construction of random cycles based the simulated system in §3.1. In the following, we will discuss the rest of the challenges in the algorithmic design and how we resolve them.

3.3.1. Simulation of the $\underline{S}$-system. We have described the simulated $\underline{S}$-system in §3.1, and the main purpose of this simulated system is to help decide triggering periods and form cycles.

An immediate important question is whether the simulated $\underline{S}$-system can be correctly simulated. Since the SCU algorithm is implemented under sales data (or censored demand), we do not know the exact demand in a period whenever a lost-sale occurs. For example, if the on-hand inventory level in our SCU-system in a period is zero, we do not know the true demand for this period since the sales is always zero regardless of demand. In this case, we cannot simulate the $\underline{S}$-system in question (as the system gives us insufficient demand information). This shows that we must design the learning algorithm in such a way that it yields the necessary demand information for simulating the $\underline{S}$-system correctly.

A sufficient condition for achieving the correct simulation of the $\underline{S}$-system is to ensure that our SCU-system always has no lower on-hand inventory than the simulated $\underline{S}$-system. To see that, suppose the states of our system and the simulated $\underline{S}$-system at the beginning of period t are $\left(q^a_{t-1}, q^a_{t-1}, \ldots, q^a_{t-L+1}, I^a_t\right)$, $a = SCU$, $\underline{S}$, respectively. Then, the on-hand inventory level at the beginning of period $t+1$ will be

$$I^a_{t+1} = q^a_{t-L+1} + \left(I^a_t - d_t\right)^+, \qquad a = SCU,\ \underline{S}.$$

In general, we may not be able to simulate the $\underline{S}$ system using only the sales quantity $\min(I_t^{SCU}, d_t)$. However, if $I_t^{SCU} \geq I_t^{\underline{S}}$, then the $\underline{S}$-system can be correctly simulated because

$$I_{t+1}^{\underline{S}} = q_{t-L+1}^{\underline{S}} + \left(I_t^{\underline{S}} - d_t\right)^+ = q_{t-L+1}^{\underline{S}} + \left[I_t^{\underline{S}} - \min\left(d_t, I_t^{SCU}\right)\right]^+.$$

This shows that, under the condition $I_t^{SCU} \geq I_t^{\underline{S}}$ for all t, the $\underline{S}$-system can be correctly simulated by pretending that the demand in period t is equal to the sales quantity in the SCU-system. This sufficient condition is will be carefully embedded in the design of our algorithm, which will be formally established in Lemma 1 in §4.2.

3.3.2. The bridging G-system, and its connection with the SCU-system. We introduce an auxiliary (non-implementable) bridging system, which we refer to as the G-system. The G-system is defined as follows: i) for cycle $k = 1, 2, \ldots$, it implements base-stock policy S_k as prescribed by the algorithm (which starts in period τ_k and ends in period $\tau_{k+1} - 1$); and ii) its state at the beginning of period 1 is set at $(S_1, 0, \ldots, 0)$, and its state at the start of cycle $k \geq 2$ (i.e., in period τ_k) is *artificially set* as

$$\left(d_{\tau_k - 1}, d_{\tau_k - 2}, \ldots, d_{\tau_k - L}, S_k - \sum_{t=\tau_k - L}^{\tau_k - 1} d_t\right).$$

Note that the main feature in the G-system is that, its inventory state at the beginning of each cycle is artificially set (hence not implementable). This change of state essentially removes the end-of-cycle effect (from the previous cycle) when implementing a different base-stock policy for the new cycle. Thus the total cost of the G-system, with the total number of cycles denoted by N, can be written as

$$\sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1} - 1). \tag{5}$$

Recall that this function G is convex with respect to the base-stock level S_k used in every cycle. It is well-known that dynamic optimization problem with a convex cost function (5) is amenable for online algorithm design (see, e.g., Hazan (2016)). Our algorithm will be based on the stochastic gradient descent method for minimizing objective function (5) of the G-system, which requires the gradient evaluation of G with respect to S_k.

However, there are still several significant challenges in evaluating the G-system based on the (censored) demand data collected from the SCU-system, due to the difference in their starting states. Our learning algorithm modifies, using historical (censored) demand information, the base-stock level from cycle to cycle. Clearly, the prescribed new base-stock level for the next cycle can be either higher or lower than the previous base-stock level, each creating critical issues. This is because, due to positive lead times, when a new base-stock level is suggested by the SCU algorithm for the following cycle, there is a random transition time before this new base-stock policy can be fully implemented (with the desired starting state).

Now, suppose that the base-stock level for period $\tau_k - 1$ is S_{k-1}, and that the SCU algorithm recommends a new base-stock policy S_k in period τ_k for the next cycle. In the following, we discuss the main issues encountered for the two cases, $S_k < S_{k-1}$ and $S_k \geq S_{k-1}$. To tackle the difficulties arising in the first case $S_k < S_{k-1}$, we shall introduce a new concept called the withheld inventory. To tackle the difficulties arising in the second case $S_k \geq S_{k-1}$, we adopt a double-phase gradient estimation approach.

3.3.3. The concept of withheld inventory. In the first case where $S_k < S_{k-1}$, the inventory position of the SCU-system in the first few periods of cycle k may be higher than S_k even if no order is placed. In this case, if we blindly and naively implement the base-stock policy S_k, we may suffer from a severe consequence that the $\underline{S}$-system may not be simulated correctly. Indeed, under this case, the desired order quantity at the beginning of cycle k may be 0 (if the inventory position after satisfying demand in period $\tau_k - 1$ is still no lower than S_k). However, ordering 0 in period τ_k will affect the on-hand inventory level of the SCU-system at the start of period $\tau_k + L$. If, for instance, the on-hand inventory level of the SCU-system at the beginning of period $\tau_k + L$ is 0, then it will reveal no demand information for period $\tau_k + L$. As a consequence, as we discussed earlier in §3.3.1, the $\underline{S}$-system cannot be simulated correctly in period $\tau_k + L$. This shows that we cannot naively follow the exact base-stock policy S_k, but need to revise the policy in such a way that the sufficient demand information for simulating the $\underline{S}$-system can be yielded.

Our approach to resolve this issue is to order the minimum but sufficient quantity to guarantee the correct simulation of the $\underline{S}$-system, but mark any excess on-hand inventory as what-we-define *withheld inventory*. (Note that the detailed formulae for the withheld inventory and its evolution are given in the description of the SCU algorithm.) At a high-level, in each period t, we shall divide the total on-hand inventory I_t^{SCU} into two parts, namely, the withheld on-hand inventory denoted by $\hat{I}_t^{SCU}$, and the regular (or non-withheld) on-hand inventory denoted by $\tilde{I}_t^{SCU}$. When making replenishment decisions in cycle k, we operate the base-stock policy S_k based on the regular inventory position only. More precisely, the order quantity is given in (4), the difference between S_k and the regular inventory position (rather than the total inventory position). Also, when satisfying demands, the withheld inventory is used only when the regular on-hand inventory has been exhausted.

The proposed (modified) base-stock policy based only on regular (or non-withheld) inventory position enables the system to *gradually* adjust its base-stock level from S_{k-1} down to S_k. This modification is essential because it ensures that the SCU system orders enough (no less than what the $\underline{S}$-system orders) in each period, in order to gather sufficient demand information that guarantees the correction simulation of the $\underline{S}$-system. Note that when all the withheld on-hand inventory is consumed by demand, the SCU-system will coincide with the G-system. The exact connection between the SCU-system with the withheld inventory and the G-system will be formally established in Lemma 2 in §4.2, which plays an essential role in comparing costs between our SCU algorithm and the optimal base-stock policy.

3.3.4. The double-phase gradient estimation. In the second case where $S_k \geq S_{k-1}$, because the G-system artificially sets its on-hand inventory level to $S_k - \sum_{t=\tau_k-L}^{\tau_k-1} d_t$ at the beginning of period τ_k, this particular on-hand inventory level could be higher than the on-hand inventory level of the SCU-system at the beginning of cycle k. At the beginning of period τ_k, the inventory vector of the SCU-system, having just experienced no lost-sales for L consecutive periods, is

$$\Big(d_{\tau_k-1}, d_{\tau_k-2}, \ldots, I_{\tau_k}^{SCU} = S_{k-1} - \sum_{t=\tau_k-L}^{\tau_k-1} d_t + \hat{I}_t\Big).$$

This is different from the starting state of the G-system, which according to our definition is

$$\left(d_{\tau_k-1}, d_{\tau_k-2}, \ldots, S_k - \sum_{t=\tau_k-L}^{\tau_k-1} d_t\right).$$

Since the on-hand inventory level of the G-system could be higher than that of the SCU-system in period τ_k, it leads to the following critical issue: Due to demand censoring and the same reasoning as in the first case, we may not able to obtain sufficient demand information from the SCU-system to compute the total cost, nor its gradient, of the G-system during periods $\tau_k, \tau_k+1, \ldots, \tau_k'-1$ with respect to the base-stock level S_k. This is precisely the reason why we need to use two phases for each cycle $k \geq 2$ in the design of our algorithm: In the triggering period τ_k', having just experienced no lost-sales for L consecutive periods, the G-system and our SCU-system become identical during the second phase if all the withheld inventory in the SCU-system is ignored.

Regardless of $S_k \leq S_{k-1}$ or $S_k \geq S_{k-1}$, we will show in Lemma 2 in §4.2 that during the second phase of each cycle, the SCU-system always has no less on-hand inventory than that of the G-system . This enables us to compute (and simulate) the total cost of the G-system during the second phase of cycle k as well as its gradient with respect to S_k. Thus, we can construct an estimate of the gradient of the entire cycle cost based on the demand data collected from the second phase. This clearly gives a *biased* estimation of the gradient. Nevertheless, we will show in Lemma 6 in §4.3 that the error of this estimation is very small in expectation and it vanishes at k grows.

3.4. Computation of Gradient $\nabla G(S,a,b)$

Let $I_t(S)$ and $q_t(S)$ denote the on-hand inventory and the ordering quantity in period t under the base-stock policy S, respectively. Also let $I_t'(S)$ and $q_t'(S)$ denote their respective gradients with respect to the base-stock level S. Since

$$\nabla G(S,a,b) = \sum_{t=a}^{b} \Big[h \cdot \mathbb{1}\left(I_t'(S)=1, D_t < I_t(S)\right) - p \cdot \mathbb{1}\left(I_t'(S)=1, D_t > I_t(S)\right) \Big],$$

we only need to keep track of $I_t(S)$ and $I_t'(S)$ from period a to b. The inventory level $I_t(S)$ is easy to compute. For $I_t'(S)$, it follows from Theorem 1 in Huh et al. (2009a) that

$$q_t'(S) = I_{t-1}'(S) \cdot \mathbb{1}\left[D_{t-1} > I_{t-1}(S)\right], \tag{6}$$

$$I_t'(S) = 1 - \textstyle\sum_{i=t-L+1}^{t} q_i'(S), \tag{7}$$

Thus, $I_t'(S)$ and $q_t'(S)$ can be computed recursively if we can evaluate $\mathbb{1}\left[D_{t-1} > I_{t-1}(S)\right]$ and have the necessary boundary conditions.

In the SCU algorithm, we need to compute the gradient $\nabla G(S_k, \tau_k', \tau_{k+1}-1)$ of the G-system (*not the SCU-system*). Note that $\nabla G(S_k, \tau_k', \tau_{k+1}-1)$ represents the partial derivative with respect to S_k assuming τ_k' and τ_{k+1} are fixed. The boundary conditions for the G-system are $(q_t^G)'(S)=0$ for $t<\tau_k'$ and $(q_{\tau_k'}^G)'(S)=1$. To evaluate $\mathbb{1}\left[D_{t-1} > I_{t-1}^G(S_k)\right]$ for $\tau_k' \le t < \tau_{k+1}-1$, we need the demand information D_{t-1} in relation to $I_{t-1}^G(S_k)$. Since the only available demand data is from the SCU-system, the comparison between D_{t-1} and $I_{t-1}^G(S_k)$ is possible only when $I_{t-1}^{SCU} \ge I_{t-1}^G$. This is true, according to the design of our algorithm and Lemma 2, for the second phase of each cycle. Thus, the gradient $\nabla G(S_k, \tau_k', \tau_{k+1}-1)$ can be readily computed.

4. Performance Analysis and Discussions

We first formally define regret. Given a sample path $\omega = \{d_1, d_2, \ldots, \}$ of the demand process, the T-period regret of the SCU algorithm is defined as the difference between the clairvoyant optimal cost (under full information) and the cost incurred by SCU over T periods. More specifically,

$$\mathcal{R}_T^{\mathbf{SCU}}(\omega) = \sum_{t=1}^{T} \left(C_t^{SCU}(\omega) - C_t^{S^*}(\omega) \right),$$

where $C_t^{SCU}(\omega)$ is the cost incurred in period t by our nonparametric (closed-loop) SCU algorithm, and $C_t^{S^*}(\omega)$ is the cost incurred in period t by the system operated under the (clairvoyant) optimal base-stock level S^*. The average regret of SCU algorithm is $\mathbb{E}\left[\mathcal{R}_T^{\mathbf{SCU}}\right]$, and the average regret per period is defined as $\mathbb{E}\left[\mathcal{R}_T^{\mathbf{SCU}}\right]/T$.

4.1. Main Result of the Paper

We formally state the main theoretical result of this paper below.

Theorem 1. *Suppose Assumption 1 holds. For each problem instance of the lost-sales inventory system with a fixed positive lead time L under censored demand information, the expected regret of the SCU algorithm is upper bounded by $O(\sqrt{T})$. That is, there exists some positive constant K, such that the expected regret of SCU algorithm satisfies*

$$\mathbb{E}\left[\mathcal{R}_T^{\mathbf{SCU}}\right] \leq K\sqrt{T}, \qquad \textit{for all} \quad T \geq 1.$$

In other words, the average regret per period approaches 0 at the rate of $O(1/\sqrt{T})$.

Our main result in Theorem 1 is significant in the sense that the newly designed simulated cycle-update algorithm improves the expected cumulative regret from $O(T^{2/3})$ to $O(T^{1/2})$, which matches the lower bound of regret for any learning algorithms (see Proposition 1 below) and resolves the open question raised by Huh et al. (2009a).

Since the constant K in the above regret bound has a complex expression, we only explicitly spell out its dependence on the basic problem primitives. The constant K is proportional to $(\max(h,p))^2\left(\bar{S}-\underline{S}\right)^2(1/c_1)^{2L}(1/c_2)^L$. Note that the term $(1/c_1)^{2L}(1/c_2)^L$ scales exponentially in L. (We remark that the constant K in the regret bound in Huh et al. (2009a) includes the term $(1/c_1)^{4L}$, which also scales exponentially in L.) This exponential dependence of the constant K on L in our regret analysis is mainly due to our theoretical upper bound on the number of periods between triggering periods, which perhaps can be further tightened. This leaves an interesting open question that whether one can prove the necessity of this exponential dependence of the constant K on L while keeping the optimal square-root rate on T. Finally, we remark that despite the relatively poor (theoretical) scaling of the constant K in L, our numerical results in §5 demonstrate that the SCU algorithm performs well even when the lead time $L = 20$.

We show that the regret rate $O(\sqrt{T})$ is tight, which is formally stated below.

Proposition 1. *Suppose $T > 5$. Even with uncensored demand, there exist problem instances such that the expected regret for any learning algorithm is lower bounded by $\Omega(\sqrt{T})$.*

The problem instance with continuous demand constructed for Proposition 1 is very similar to the discrete demand example constructed by Besbes and Muharremoglu (2013). Following their arguments, we provide the proof of Proposition 1 in the Appendix, for the sake of completeness.

4.2. Building Blocks for Regret Analysis

To prove our main result (i.e., Theorem 1), we first need to establish several important building blocks for the regret analysis, which are presented below. Their detailed proofs are given in the Appendix.

The first result ensures that the cycles used in designing the SCU algorithm are well defined: By maintaining no less on-hand inventory in SCU-System than in the $\underline{S}$-system, the system dynamics of the $\underline{S}$-system can always be correctly simulated.

Lemma 1. *The SCU-system always has no less on-hand inventory than the simulated $\underline{S}$-system.*

The next result ensures that the gradient of the G-system, which is used in the SCU algorithm, can indeed be computed using (censored) demand data collected from the SCU-system. Recall that only the gradient in the second phase of each cycle $k \geq 2$ is computed and used in the SCU algorithm.

Lemma 2. *For the SCU algorithm, in each period of the second phase of any cycle $k \geq 2$, the SCU-system has no less on-hand inventory than the G-system.*

The following two lemmas delineate the relationships between demand characteristics and lost-sales events in the lost-sales inventory system, and they will play important roles in the proof of our main result. They also explain why Assumption 1 is needed for the main result to hold.

Lemma 3. *For the simulated $\underline{S}$-system, if $d_k \leq \frac{\underline{S}}{L+1}$ for consecutive $2L$ periods $k = t$ to $t+2L-1$, then there is no lost-sales in the simulated $\underline{S}$-system from period $t+L$ to period $t+2L-1$.*

Lemma 4. *For a base-stock system with base-stock level S, if $d_k > \frac{S+\delta}{L+1}$ for $k = t, \ldots, t+L$, then the total lost-sales amount from period t to period $t+L$ is at least δ.*

Following Lemmas 3 and 4, we define, for any period t, two random variables:

$$\underline{t} = \min_k \left\{ k \geq t+2L-1 : \max_{k-2L+1 \leq i \leq k} d_i \leq \frac{\underline{S}}{L+1} \right\}, \tag{8}$$

$$\bar{t} = \min_k \left\{ k \geq t+L-1 : \min_{k-L+1 \leq i \leq k} d_i \geq \frac{S^*+1}{L+1} \right\}. \tag{9}$$

By Assumption 1, both $\underline{t}$ and $\bar{t}$ are well-defined. In fact, $\underline{t} - t$ and $\bar{t} - t$ are known as geometric random variables of orders $2L$ and L, and parameters $1/c_1$ and $1/c_2$, respectively (see Philippou et al. (1983), Proposition 2.2). They represent the number of periods it takes after t such that demand is no more than $\underline{S}/(L+1)$ for $2L$ consecutive periods for the first time, and no less than $S^* + 1/(L+1)$ for L consecutive periods for the first time, respectively. By Lemma 3, between t and $\underline{t}$, there must exist L consecutive periods such that the $\underline{S}$-system has no lost-sales. Similarly, by Lemma 4, the S^*-system must have at least one unit of lost-sales between t and $\bar{t}$.

The following lemma discusses the impact of perturbing the initial inventory vector in an inventory system that implements a base-stock policy, and it will be used in comparing the SCU- and G- systems during a cycle. It states that the perturbation does not amplify during the cycle.

LEMMA 5. *Fix a sample path of demand process and consider two systems, referred to as the original system and β-system respectively, both operating the same base-stock policy S, but their states at the beginning of the first period are $[q_1, q_0, \dots, q_{2-L}, I_1]$ and $[q_1+\beta, q_0, \dots, q_{2-L}, I_1-\beta]$, with $0 \le \beta \le I_1$. Then, we have $\left|I_t^o - I_t^\beta\right| \le \beta$ for all $t \ge 1$, where I_t^o and I_t^β are the on-hand inventory levels of the original system and the β-system, respectively.*

4.3. Proof of the Main Result

In what follows, we prove Theorem 1 based on Lemmas 1–5 established above. The proof makes use of three bridging systems, and we will show that the cost difference between any two adjacent systems is on the order $O(\sqrt{T})$.

The three bridging systems are the $\underline{S}$-system, the G-system, and the $\overline{\text{SCU}}$-system which is defined as the SCU-system but with holding costs for the withheld on-hand inventory setting to 0. Figure 2 shows the roadmap of our proof.

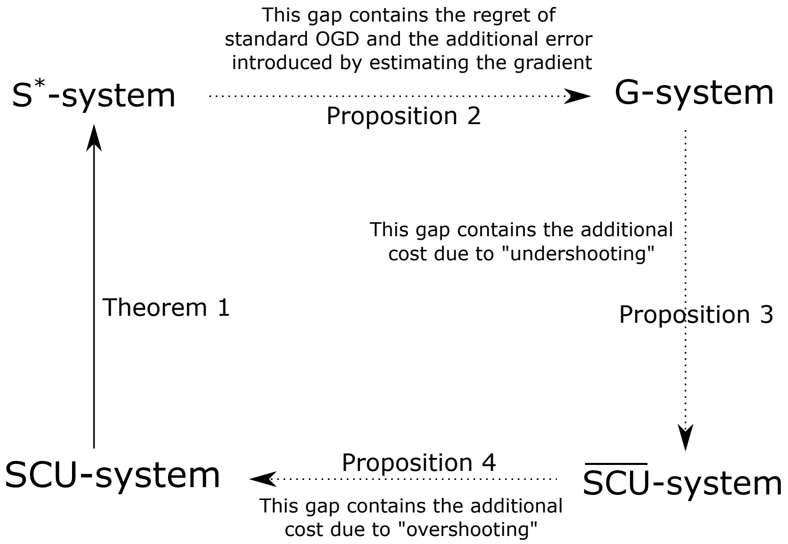
This gap contains the regret of standard OGD and the additional error introduced by estimating the gradient
S*-system
G-system
Proposition 2
This gap contains the additional cost due to "undershooting"
Theorem 1
Proposition 3
SCU-system
Proposition 4
$\overline{\text{SCU}}$-system
This gap contains the additional cost due to "overshooting"

Figure 2 A roadmap for the proof of Theorem 1

For a fixed T, we let N denote the number of cycles (including the last possibly incomplete one). With a slight abuse of notation, we let $\tau_{N+1} = T+1$ so the last cycle ends at T. Note that N is a random variable that depends on the demand process.

In the following, we first show that the difference between the expected costs of S^*-system and G-system is upper bounded by $O(\sqrt{T})$.

PROPOSITION 2. *There exists some positive constant K_1, such that*

$$\mathbb{E}\left[\sum_{k=1}^{N} G(S_k,\tau_k,\tau_{k+1}-1)\right]-\mathbb{E}\left[\sum_{t=1}^{T} C_t^{S^*}\right]\le K_1\sqrt{T}.$$

Proof. For the first cycle, we have

$$\begin{aligned}\mathbb{E}\left[G(S_1,\tau_1,\tau_2-1)\right]-\mathbb{E}\left[\sum_{t=\tau_1}^{\tau_2-1} C_t^{S^*}\right] &\le \mathbb{E}\left[\tau_2-\tau_1\right]\cdot\max(h,p)\left(\bar{S}-\underline{S}\right)\\ &\le \mathbb{E}\left[\underline{t}-\tau_1\right]\cdot\max(h,p)\left(\bar{S}-\underline{S}\right)\\ &=\frac{1-c_1^{2L}}{(1-c_1)c_1^{2L}}\cdot\max(h,p)\left(\bar{S}-\underline{S}\right),\end{aligned}$$

where the last equality follows from Proposition 2.1 of Philippou et al. (1983). Similarly, for the last cycle, we have

$$\mathbb{E}\left[G(S_N,\tau_N,T)\right]-\mathbb{E}\left[\sum_{t=\tau_N}^{T} C_t^{S^*}\right]\le 2\cdot\frac{1-c_1^{2L}}{(1-c_1)c_1^{2L}}\cdot\max(h,p)\left(\bar{S}-\underline{S}\right).$$

Every cycle $1<k<N$ contains two phases. The first phase is from period τ_k to period $\tau_k'-1$, and the second phase is from period τ_k' to period $\tau_{k+1}-1$. Recall that the cost gradient for the first phase cannot be evaluated due to lack of demand information. To complete the proof of Proposition 2, we need the following result which shows that, the gradient for the first phase approaches that of the second phase in expectation, which can be evaluated, when k increases. The proof of this technical lemma is based on uniform ergodicity, which is deferred to the Appendix.

LEMMA 6. *For $1<k<N$, we have*

$$\left|\mathbb{E}\left[\nabla G(S_k,\tau_k,\tau_k'-1)\right]-\mathbb{E}\left[\nabla G(S_k,\tau_k',\tau_{k+1}-1)\right]\right|=o\left(1/\sqrt{k}\right). \tag{10}$$

The above result shows that, although a biased gradient is used in the SCU algorithm in the search for the best base-stock level, it is close to the true gradient when k is large and converges at a rate faster than $o(1/\sqrt{k})$. Applying this result, we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{N} G(S_i,\tau_k,\tau_{k+1}-1)-\sum_{t=1}^{T} C_t^{S^*}\right]\\ &=\mathbb{E}\left[\sum_{k=1}^{N}\left(G(S_k,\tau_k,\tau_{k+1}-1)-G(S^*,\tau_k,\tau_{k+1}-1)\right)\right]\end{aligned}$$

$$\begin{aligned}
&\le \mathbb{E}\left[\sum_{k=1}^{N} \nabla G(S_k, \tau_k, \tau_{k+1}-1)(S_k - S^*)\right] \\
&\le \mathbb{E}\left[\sum_{k=2}^{N-1} \nabla G(S_k, \tau_k, \tau_{k+1}-1)(S_k - S^*)\right] + 3 \cdot \frac{1-c_1^{2L}}{(1-c_1)c_1^{2L}} \cdot \max(h,p)(\bar{S} - \underline{S}) \\
&\le \mathbb{E}\left[\sum_{k=2}^{N-1} \left(2\nabla G(S_k, \tau_k', \tau_{k+1}-1)(S_i - S^*) + o(1/\sqrt{k})\right)\right] + 3 \cdot \frac{1-c_1^{2L}}{(1-c_1)c_1^{2L}} \cdot \max(h,p)(\bar{S} - \underline{S}) \\
&\le \mathbb{E}\left[\sum_{k=2}^{N-1} \left(\frac{\sqrt{k}}{2\gamma}\left((S_k - S^*)^2 - (S_{k+1} - S^*)^2\right)\right)\right] + \mathbb{E}\left[\sum_{k=2}^{N-1} \frac{2\gamma \nabla G(S_k, \tau_k', \tau_{k+1}-1)^2}{\sqrt{k}}\right] + o\left(\sqrt{T}\right),
\end{aligned} \tag{11}$$

where the first inequality follows from the convexity of function $G(S, \tau_k, \tau_{k+1}-1)$ in S, the third inequality is due to (10), and the last inequality is because of, by our SCU algorithm,

$$(S_{k+1} - S^*)^2 \le (S_k - S^*)^2 - \frac{4\gamma}{\sqrt{k}}(S_k - S^*)\nabla G(S_k, \tau_k', \tau_{k+1}-1) + \frac{4\gamma \left(\nabla G(S_k, \tau_k', \tau_{k+1}-1)\right)^2}{k}.$$

We evaluate the first term on the right hand side of (11) as follows:

$$\begin{aligned}
&\mathbb{E}\sum_{k=2}^{N-1}\left[\frac{\sqrt{k}}{2\gamma}\left((S_k - S^*)^2 - (S_{k+1} - S^*)^2\right)\right] \\
&\le \frac{1}{\gamma}\mathbb{E}\left[\frac{\sqrt{2}}{2}(S_2 - S^*)^2 - \frac{\sqrt{N-1}}{2}(S_N - S^*)^2\right] + \frac{1}{2\gamma}\mathbb{E}\sum_{k=3}^{N-1}\left[(\sqrt{k} - \sqrt{k-1})(S_k - S^*)^2\right] \\
&\le \frac{\sqrt{2}}{2\gamma}(\bar{S} - \underline{S})^2 + \frac{1}{2\gamma}\mathbb{E}\left[\sum_{k=3}^{T}(\sqrt{k} - \sqrt{k-1})(\bar{S} - \underline{S})^2\right] \\
&= \frac{\sqrt{T}}{2\gamma}(\bar{S} - \underline{S})^2.
\end{aligned} \tag{12}$$

To evaluate the second term on the right hand side of (11), we first focus on the term $\mathbb{E}[(\nabla G(S_k, \tau_k', \tau_{k+1}-1))^2]$. From Lemma 3, $\tau_{k+1}-1$ is no larger than $\underline{\tau_k'}$ with probability 1. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\nabla G(S_k, \tau_k', \tau_{k+1}-1)\right]^2 &\le \left[\max(h,p)\left(\bar{S} - \underline{S}\right)\right]^2 \cdot \mathbb{E}\left[\left(\underline{\tau_k'} - \tau_k'\right)^2\right] \qquad (13) \\
&= \left[\max(h,p)\left(\bar{S} - \underline{S}\right)\right]^2 \cdot \frac{2 + (4L-1)c_1^{2L} - (4L+1)c_1^{2L+1} + c_1^{4L} - c_1^{4L+1}}{c_1^{4L} - c_1^{4L+2}},
\end{aligned}$$

where the equality above follows from Proposition 2.1 in Philippou et al. (1983). Thus, we obtain, for some constant K_1,

$$\mathbb{E}\left[\sum_{k=2}^{N} \frac{2\gamma \nabla G(S_k, \tau_k', \tau_{k+1}-1)^2}{\sqrt{k}}\right] \le \mathbb{E}\left[\sum_{k=1}^{T} \frac{2\gamma \nabla G(S_k, \tau_k', \tau_{k+1}-1)^2}{\sqrt{k}}\right] \le K_1 \cdot \sqrt{T}. \tag{14}$$

Combining (11), (13) and (14), we complete the proof of Proposition 2. **Q.E.D.**

Because function G is a convex function with respect to the base-stock level and is minimized at S^* in expectation, we can see from Proposition 2 that S_k is converging to the optimal base-stock level S^*. As a by-product, we obtain the following result in Lemma 7, which will be used in the proof of Proposition 4. The detailed proof of this lemma is given in the Appendix.

LEMMA 7. *There exists some positive constant K_2, such that*

$$\mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}(S_k > S^* + 1/2)\right] \leq K_2\sqrt{T}.$$

We next compare the G-system with the $\overline{\text{SCU}}$-system. The difference between these two systems lies in the "undershooting" of the $\overline{\text{SCU}}$-system. That is, both systems operate under the same base-stock level, but at the beginning of each cycle, the $\overline{\text{SCU}}$-system potentially has less on-hand inventory and has to order more to keep the same inventory position as G-system. We will show that the cost difference created by "undershooting" the target levels is upper bounded by $O(\sqrt{T})$ in expectation.

PROPOSITION 3. *There exists some positive constant K_3, such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\overline{SCU}}\right] - \mathbb{E}\left[\sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1} - 1)\right] \leq K_3\sqrt{T}.$$

Proof. Let $\tilde{I}_t^{SCU}$and I_t^G denote the on-hand inventory levels of $\overline{\text{SCU}}$- and G- systems, respectively. Then, for every sample path, we have

$$\begin{aligned}\sum_{t=1}^{T} C_t^{\overline{\text{SCU}}} - \sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1} - 1) &\leq \sum_{t=1}^{T} \max(h,p)\left|I_t^G - \tilde{I}_t^{SCU}\right| \\ &= \sum_{k=1}^{N} \sum_{t=\tau_i}^{\tau_{k+1}-1} \max(h,p)\left|I_t^G - \tilde{I}_t^{SCU}\right|. \qquad (15)\end{aligned}$$

For the first cycle, we have $I_t^G = \tilde{I}_t^{SCU}$ for every period t. For cycle $k \geq 2$, if $S_k < S_{k-1}$, then by the construction of the SCU algorithm, $I_t^G = \tilde{I}_t^{SCU}$ for every period t in cycle k; if $S_k \geq S_{k-1}$, then I_t^G may differ from $\tilde{I}_t^{SCU}$ for t in the first phase of the cycle (i.e., periods from τ_k to $\tau_k' - 1$), and they will become the same from τ_k' until $\tau_{k+1} - 1$.

Suppose $S_k \geq S_{k-1}$, we will show that

$$\left|I_t^G - \tilde{I}_t^{SCU}\right| \leq I_{\tau_k}^G - \tilde{I}_{\tau_k}^{SCU} \leq S_k - S_{k-1}, \text{ for periods } t = \tau_k, \ldots, \tau_k' - 1.$$

We first prove the second inequality $I^G_{\tau_k} - \tilde{I}^{SCU}_{\tau_k} \leq S_k - S_{k-1}$. For the G-system, its inventory vector at period τ_k is $\left[d_{\tau_k-1}, d_{\tau_k-2}, \ldots, d_{\tau_k-L}, S_k - \sum_{i=\tau_k-L}^{\tau_k-1} d_i\right]$. For the SCU-system, if $\hat{I}_{\tau_k-1} = 0$, then the inventory vector at τ_k would be $\left[d_{\tau_k-1} - S_{k-1} + S_k, d_{\tau_k-2}, \ldots, d_{\tau_k-L}, S_{k-1} - \sum_{i=\tau_k-L}^{\tau_k-1} d_i\right]$. In this case we have

$$I^G_{\tau_k} - \tilde{I}^{SCU}_{\tau_k} = I^G_{\tau_k} - I^{SCU}_{\tau_k} = S_k - S_{k-1}.$$

If $\hat{I}_{\tau_k-1} > 0$, then some of the withheld on-hand inventory will be included back to the regular on-hand inventory by equation $\hat{I}_{\tau_k} = (\hat{I}_{\tau_k} - (S_k - S_{k-1}))^+$, and as a result, the regular on-hand inventory in the SCU-system may be higher and we will have $I^G_{\tau_k} - \tilde{I}^{SCU}_{\tau_k} < S_k - S_{k-1}$. Hence in all cases $I^G_{\tau_k} - \tilde{I}^{SCU}_{\tau_k} \leq S_k - S_{k-1}$ is satisfied. Then, we apply Lemma 5 to obtain $\left|I^G_t - \tilde{I}^{SCU}_t\right| \leq I^G_{\tau_k} - \tilde{I}^{SCU}_{\tau_k}$ for all $t = \tau_k, \ldots, \tau'_k$. Combining the two scenarios shows $\left|I^G_t - \tilde{I}^{SCU}_t\right| \leq |S_k - S_{k-1}|$ for $t = \tau_k, \ldots, \tau'_k - 1$ and $\left|I^G_t - \tilde{I}^{SCU}_t\right| = 0$ for $t = \tau'_k, \ldots, \tau_{k+1} - 1$.

Taking expectation on both sides of (15), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\overline{\text{SCU}}} - \sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1} - 1)\right] \leq \max(h,p)\mathbb{E}\left[\sum_{k=1}^{N}\sum_{t=\tau_k}^{\tau'_k}\left|I^G_t - \tilde{I}^{SCU}_t\right|\right] \\
&\leq \max(h,p)\mathbb{E}\left[\sum_{k=2}^{N}(\tau'_k - \tau_k + 1)\,|S_k - S_{k-1}|\right] \leq \max(h,p)\mathbb{E}\left[\sum_{k=2}^{T}(\tau'_k - \tau_k + 1)\,|S_k - S_{k-1}|\right] \\
&= \max(h,p)\mathbb{E}\left[\sum_{k=2}^{T}|S_k - S_{k-1}|\right]\mathbb{E}\left[(\tau'_k - \tau_k + 1)\right] = \max(h,p)\mathbb{E}\left[\sum_{k=2}^{N}|S_k - S_{k-1}|\right]\frac{1 - c_1^{2L}}{(1 - c_1)c_1^{2L}} \\
&\leq \max(h,p)^2\mathbb{E}\left[\sum_{k=2}^{N}\frac{2\gamma}{\sqrt{k}}(\tau'_{k-1} - \tau_{k-1})\right]\frac{1 - c_1^{2L}}{(1 - c_1)c_1^{2L}} \leq \left(\max(h,p)\frac{1 - c_1^{2L}}{(1 - c_1)c_1^{2L}}\right)^2\sum_{k=1}^{T}\frac{2\gamma}{\sqrt{k}} \leq K_3 \cdot \sqrt{T}
\end{aligned}$$

for some constant K_3. The first equality above is by the independence of $|S_k - S_{k-1}|$ and $\tau_k - \tau_k$, and the second equality and the inequality after that are by Proposition 2.1 in Philippou et al. (1983). This completes the proof of Proposition 3. **Q.E.D.**

Following the roadmap in Figure 2, the last part of the regret analysis is to bound the gap between the SCU-system and the $\overline{\text{SCU}}$-system. The cost difference between these two systems is upper bounded by the total holding cost of the withheld on-hand inventory. The following lemma shows that this part is also bounded by $O(\sqrt{T})$ in expectation.

Proposition 4. *There exists some positive constant K_4 such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} C_t^{SCU}\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{\overline{SCU}}\right] \leq K_4\sqrt{T}.$$

Proof. First, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} C_t^{SCU}\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{\overline{SCU}}\right] \leq h\sum_{t=1}^{T}\mathbb{E}\left[\hat{I}_t\right]. \tag{16}$$

According to the SCU algorithm, we have $\hat{I}_{t+1} = (\hat{I}_t - (d_t - \tilde{I}_t)^+)^+$ from period to period during the cycle, thus the withheld inventory is gradually consumed by demand; and when going from one cycle to the next, new withheld inventory may be created by $\hat{I}_{\tau_{k+1}} := (\hat{I}_{\tau_{k+1}} - (S_{k+1} - S_k))^+$ at the beginning of cycle $k+1$. This shows that, the only source of new withheld inventory is generated at the beginning of the cycles, and the maximum added at the beginning of cycle $k+1$ is $(S_k - S_{k+1})^+$, $k = 1, 2, \ldots, N$.

When evaluating the RHS of (16), instead of evaluating it vertically by finding the total amount of $\hat{I}_t$ in every period and then adding up over periods, we compute it horizontally by identifying the total number of periods in which a withheld inventory unit stays in the system and then adding up over all the withheld inventory units. At the beginning of period τ_{k+1}, a maximum of $(S_k - S_{k+1})^+$ units of the withheld inventory are created in the system, and from $\hat{I}_{t+1} = (\hat{I}_t - (d_t - \tilde{I}_t)^+)^+$, it is seen that within a cycle, the amount of the withheld on-hand inventory is non-increasing.

In the following, we consider three cases, namely, 1) $S_k > S^* + 1/2$ and $S_{k+1} \leq S^* + 1/2$, and 2) $S_k > S^* + 1/2$ and $S_{k+1} > S^* + 1/2$, and 3) $S_k \leq S^* + 1/2$,

For the first case, when there are L consecutive periods of demands higher than $\frac{S^*+1}{L+1}$, we know by Lemma 4 that there would be at least 1/2 lost-sales in the system with base-stock level S_{k+1} by the time $\bar{\tau}_{k+1}$, if there is no withheld inventory. This also implies that if there is the additional withheld inventory introduced at τ_{k+1}, then it will be reduced by at least 1/2 by the time $\bar{\tau}_{k+1}$, since the withheld inventory must be consumed by demand before any lost-sales can occur. We further denote $\bar{\bar{\tau}}_{k+1}$ as the period after τ_{k+1} such that the event that demand is higher than $\frac{S^*+1}{L+1}$ for L consecutive periods happens $2(\bar{S} - S^*)$ times. Then it is clear that all the withheld inventory introduced at and before τ_{k+1} will disappear by the time $\bar{\bar{\tau}}_{k+1}$.

For the second case, we simply keep the withheld inventory introduced at τ_{k+1}, at most $\bar{S} - S_{k+1}$ units, in the system until the start of the next cycle τ_{k+1}.

For the third case, the additional withheld inventory introduced at τ_{k+1} will be consumed by demand by the time $\bar{\tau}_{k+1}$. This is because if the additional withheld inventory is positive, we must have $S_{k+1} < S_k \leq S^* + 1/2$. When there are L consecutive periods of demands higher than $\frac{S^*+1}{L+1}$, then by Lemma 4 there would be at least 1/2 lost-sales in the system with any base-stock level less than $S^* + 1/2$. Since both S_k and S_{k+1} are less than $S^* + 1/2$, the system will experience lost-sales, and therefore that the additional withheld inventory will disappear by the time $\bar{\tau}_{k+1}$.

Following the above three cases, we have

$$
\begin{aligned}
\mathbb{E}\left[h\sum_{t=1}^{T}\hat{I}_t\right] \le\ & h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2, S_{k+1}\le S^*+1/2)(\bar{S}-S_{k+1})^+\left(\bar{\bar{\tau}}_{k+1}-\tau_{k+1}\right)\right]\\
&+h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2, S_{k+1}> S^*+1/2)(\bar{S}-S_{k+1})^+\left(\tau_{k+2}-\tau_{k+1}\right)\right]\\
&+h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k \le S^*+1/2)(S_k-S_{k+1})^+\left(\bar{\tau}_{k+1}-\tau_{k+1}\right)\right]\\
\le\ & h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2)(\bar{S}-S_{k+1})^+\left[\left(\bar{\bar{\tau}}_{k+1}-\tau_{k+1}\right)+\left(\tau_{k+2}-\tau_{k+1}\right)\right]\right]\\
&+h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k \le S^*+1/2)(S_k-S_{k+1})^+\left(\bar{\tau}_{k+1}-\tau_{k+1}\right)\right]. \qquad (17)
\end{aligned}
$$

First, we bound the first term on the right hand side of (17). For some constant K_5,

$$
\begin{aligned}
& h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2)(\bar{S}-S_{k+1})^+\left[\left(\bar{\bar{\tau}}_{k+1}-\tau_{k+1}\right)+\left(\tau_{k+2}-\tau_{k+1}\right)\right]\right]\\
\le\ & h\bar{S}\cdot 2\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2)\left[\left(\bar{\bar{\tau}}_{k+1}-\tau_{k+1}\right)+\left(\tau_{k+2}-\tau_{k+1}\right)\right]\right]\\
\le\ & h\bar{S}\cdot 2\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k > S^*+1/2)\right]\cdot\left(2\left(\bar{S}-S^*\right)\mathbb{E}\left[\left(\bar{\tau}_{k+1}-\tau_{k+1}\right)\right]+2\mathbb{E}\left[\left(\underline{\tau}_{k+1}-\tau_{k+1}\right)\right]\right)\\
\le\ & h\bar{S}\cdot 2K_2\sqrt{T}\cdot\left[2(\bar{S}-S^*)\cdot\frac{1-c_2^L}{(1-c_2)c_2^L}+2\frac{1-c_1^{2L}}{(1-c_1)c_1^{2L}}\right]\le K_5\cdot\sqrt{T},
\end{aligned}
$$

where the second inequality follows from the definition of $\bar{\bar{\tau}}_{k+1}$, Lemma 3 and the definition of $\underline{t}$. The third inequality follows from Lemma 7 and by Proposition 2.1 in Philippou et al. (1983).

Next, we bound the second term on the right hand side of (17). For some constant K_6,

$$
\begin{aligned}
& h\mathbb{E}\left[\sum_{k=1}^{N-1}\mathbb{1}(S_k \le S^*+1/2)(S_k-S_{k+1})^+\left(\bar{\tau}_{k+1}-\tau_{k+1}\right)\right]\\
\le\ & h\mathbb{E}\left[\sum_{k=2}^{N}\left|S_{k-1}-S_k\right|\left(\bar{\tau}_k-\tau_k\right)\right]\\
=\ & h\mathbb{E}\left[\sum_{k=2}^{N}\left|S_{k-1}-S_k\right|\right]\cdot\mathbb{E}\left[\left(\bar{\tau}_k-\tau_k\right)\right]\\
\le\ & h\mathbb{E}\left[\sum_{k=2}^{N}\left|S_{k-1}-S_k\right|\right]\cdot\frac{1-c_2^L}{(1-c_2)c_2^L}\\
\le\ & h\max(h,p)\sum_{k=2}^{N}\frac{2\gamma}{\sqrt{k}}\mathbb{E}\left[\tau_k-\tau_{k-1}\right]\cdot\frac{1-c_2^L}{(1-c_2)c_2^L}
\end{aligned}
$$

$$
\begin{aligned}
&\le \sum_{k=1}^{T} \frac{2\gamma}{\sqrt{k}} \left[h \cdot \max(h,p) \cdot \frac{2(1-c_1^{2L})}{(1-c_1)c_1^{2L}} \cdot \frac{1-c_2^L}{(1-c_2)c_2^L} \right] \\
&\le K_6 \cdot \sqrt{T},
\end{aligned}
$$

where the first equality holds by independence, and the second inequality is by Proposition 2.1 in Philippou et al. (1983).

Combining the above two terms and (17), we have, for some constant K_3, $\mathbb{E}\left[h\sum_{t=1}^{T}\hat{I}_t\right] \le K_3 \cdot \sqrt{T}$. Then Proposition 4 follows immediately from (16). **Q.E.D.**

Combining Propositions 2, 3 and 4, we complete the proof of Theorem 1.

4.4. The SCU Algorithm for Uncensored Demand

The censored demand assumption in this paper brings two main challenges in the development of our learning algorithm. The first challenge is to guarantee the correct simulation of the $\underline{S}$-system, which requires our system to always have no less on-hand inventory than the $\underline{S}$-system. To achieve that, we have to dynamically modify the target base-stock levels. This is the reason for introducing the concept of withheld on-hand inventory. When demand is uncensored, this is not necessary, as the $\underline{S}$-system can always be simulated in each and every period. Thus, we can simply apply the target base-stock level S_k for every cycle k. The second challenge is the evaluation of gradient for our learning algorithm. With censored demand, we cannot evaluate the gradient of function G when the SCU-system less on-hand inventory than the G-system. To overcome this issue, we design two phases in each cycle $k \ge 2$ where the gradient of G-system can always be evaluated in the second phase, and that is then used to estimate the gradient for cycle k. (We double the gradient of the second phase to provide a close yet biased estimate for the gradient of the whole cycle.) With uncensored demand, this is again not necessary, as the gradient of G can always be computed.

These observations lead to a much simpler SCU algorithm for the case with uncensored demand in which neither the withheld inventory nor an additional phase for the learning cycle is needed. Denote this modified SCU algorithm for the uncensored demand case by SCU-UN, and we formally present it below. In this algorithm, the gradient $\nabla G(S_k, \tau_k, \tau_{k+1}-1)$ for cycle k is computed following the same procedure in §3.2 but using uncensored demand data in cycle k.

Algorithm 2: SCU-UN

Step 0 (**Initialization**): Start with an arbitrary target base-stock level $S_1 \in [\underline{S}, \bar{S}]$. Set $\tau_1 = 1$, cycle number $k = 1$. Let step size $\eta_k = \frac{\gamma}{\sqrt{k}}$ for $k = 1, 2, \ldots$, for some positive constant γ. Set the consecutive no lost-sales indicator $\psi := 0$. Set $t = 1$, and the initial inventory of SCU-UN-system and simulated $\underline{S}$-system to $\mathbf{x}_1^{SCU-UN} = \mathbf{x}_1^{\underline{S}} = \mathbf{0}$.

Step 1: In each period t, do the following:

(a) For SCU-UN-system, order $q_t^{SCU\text{-}UN} = \left(S_k - \sum \mathbf{x}_t^{SCU\text{-}UN}\right)^+$; for the simulated $\underline{S}$-system, order $q_t^{\underline{S}} = (\underline{S} - \sum \mathbf{x}_t^{\underline{S}})^+$.

(b) Observe demand d_t, and update the states of the SCU-UN-system and the $\underline{S}$-system according to the system dynamics (1).

(c) If there is no lost-sales in the $\underline{S}$-system, then set $\psi := \psi + 1$. Otherwise set $\psi := 0$.

(d) If $\psi = L$, then label period $t+1$ as a triggering period. Set $\tau_{k+1} = t+1$, and $\psi := 0$. Update the target base-stock level for the next cycle as

$$S_{k+1} = \mathbf{P}_{[\underline{S},\bar{S}]} \left(S_k - \eta_k \nabla G\left(S_k, \tau_k, \tau_{k+1} - 1\right)\right).$$

Set $t := t+1$ and $k := k+1$, and repeat Step 1.

This concludes the description of the SCU-UN algorithm.

Although the SCU-UN algorithm is much simpler than the SCU algorithm, the essential idea of (random) cycle-updating rule based on the simulated $\underline{S}$-system remains the same. The performance analysis of SCU-UN is similar and simpler compared to SCU, and under the same Assumption 1, it achieves a regret of rate $O(\sqrt{T})$ for the uncensored demand case. We omit the details of performance analysis for the SCU-UN algorithm.

One purpose of introducing the SCU-UN algorithm is to study the value of observing lost-demand information. That is, how much cost savings can be resulted from knowing the lost-demand information? In §5, we will conduct a numerical study of both SCU and SCU-UN to investigate the performance gap between these two cases.

4.5. Connection with Prior Literature

As we discussed in §1, our algorithms are tightly connected to the algorithms proposed in Huh et al. (2009a) and Huh and Rusmevichientong (2009). Huh et al. (2009a) considered the same problem, i.e., the lost-sales inventory system with positive lead times. Both the SCU algorithm and the algorithm proposed in Huh et al. (2009a) are stochastic gradient descent type algorithms. Both leverage the gradient information of a cycle to update the base-stock level. The major difference is as follows. The algorithm in Huh et al. (2009a) uses *fixed* and *pre-determined* cycles with *increasing* lengths and only uses the cost gradient of a single (the last) period of each cycle to carry out the updating. In contrast, the SCU algorithm is based on *a priori random* cycles that are triggered by lost-sales events, and it uses the cost gradient of the entire cycle (or at least half of it) to carry out the updating. This idea is crucial for the improved performance of the SCU algorithm, because for the existing algorithm with pre-determined cycle lengths to converge to the optimal base-stock level, the cycle length has to be increasing over time, which leads to a gap of regret rate in Huh et al. (2009a). Our regret analysis is also very different than the one used in Huh et al. (2009a), and it involves several significant new ideas including the withheld on-hand inventory and the double-phase cycle gradient estimation.

Huh and Rusmevichientong (2009) considered a much simpler lost-sales inventory system with zero lead times. They introduced a stochastic gradient descent type algorithm that updates its target inventory level in every period. When the lead time is zero, we can let each period be a cycle in the SCU algorithm and a single phase (in this case a single period) is sufficient to observe the gradient information in order to update the target inventory level. Hence, the SCU algorithm essentially reduces to the algorithm in Huh and Rusmevichientong (2009) when lead time is zero. We also remark that the algorithm in Huh et al. (2009a) does not enjoy the same reduction since their cycle lengths are increasing.

5. Computational Experiments

We conduct comprehensive numerical experiments to study on the empirical performance of the proposed SCU algorithm. We first test our algorithm against the algorithm proposed in Huh et al. (2009a) (denoted by HJMR for short), and then against the SCU-UN algorithm described in §4.4, which will reveal the value of censored demand information. The performance is evaluated by the percentage of increase in total cost of a given learning algorithm π (over the planning horizon) compared with that of the clairvoyant optimal base-stock policy. That is, we measure

$$\kappa(\pi) = \frac{\mathbb{E}[\mathcal{R}_t^\pi]}{\mathbb{E}\left[\sum_{t=1}^T C_t^{S^*}\right]} \times 100\%,$$

where S^* is the clairvoyant optimal base-stock level.

Design of experiments. We first present the design of our numerical experiments. Four lead times $L \in \{5, 10, 15, 20\}$ are considered. For the cost structure, we normalize the per-unit holding cost to $h = 1$, and consider three per-unit lost-sales costs $p \in \{50, 75, 100\}$. We consider three demand distributions: a) Gamma distribution with mean 10 and three different shape parameters $\alpha \in \{3, 5, 7\}$, b) Uniform distribution between $[0, 20]$, c) Poisson distribution with mean 10. We set $\underline{S} = 9 \cdot L + 1$ and $\bar{S} = 20 \cdot L + 1$ for the SCU, HJMR, and SCU-UN algorithms. The clairvoyant optimal solution is always contained in $[\underline{S}, \bar{S}]$. We consider four planning horizons $T \in$

$\{100, 200, 1000, 2000\}$. All systems start empty with zero initial inventory. For each testing instance, we generate 5000 sample paths of the random demand process, and use that to compute the average cost of a given learning algorithm.

Recall that the step-size $\eta_k = \gamma/\sqrt{k}$ in the proposed SCU and SCU-UN algorithms. We need to choose an appropriate γ in our numerical experiments. Since both algorithms use a cycle cost gradient (instead of a per-period cost gradient), a smaller value of γ is preferred so as to smoothen the gradient update, especially with a larger lead time L. We find that $\gamma = O(1/L)$ generally works well computationally, and in our simulation study, we fix $\gamma = 1/(4L)$ for all the instances. For HJMR, the (theoretical) γ proposed in their paper seems a little too large in computational experiments, yielding a very slow convergence. To keep the comparison fairer, we use $k\gamma$ for different values of k in HJMR, and find $k = 1/2$ gives a much better performance.

Numerical results. Tables 1 and 2 summarize the performance of the SCU, HJMR and SCU-UN algorithms under all the tested instances. We first compare the empirical performance between SCU and HJMR. Our numerical results show that HJMR converges faster in most cases (except the case of Poisson demand) when $L = 5$. The performance of both algorithms is comparable when $L = 10$. However, when $L = 15$ and $L = 20$, the SCU algorithm converges consistently faster in all cases. It seems that SCU is more robust with large lead times. Also, for both algorithms, the convergence generally becomes slower as p increases, which is consistent with the theoretical regret upper bound. It is also interesting to note that for both algorithms, we observe that the convergence may not take place within the first 200 periods (i.e., the regret might go up), which is mainly due to the fact the system with positive lead times generally takes time to stabilize.

We then compare the empirical performance between SCU and SCU-UN. As expected, SCU-UN indeed performs better in all cases. On average, the SCU-UN saves around 30% of the total regret at $T = 5000$, which demonstrates that the value of uncensored demand information is quite significant. Thus, if feasible, it pays for the inventory manager to invest in necessary information systems (e.g., online view/demand tracking systems) to capture such lost-sales information.

For the proposed SCU and SCU-UN algorithms, apart from the expected average regret, another important question is how frequent the algorithm makes an update on the target inventory position, which is determined by what-we-call triggering period, i.e., the number of periods between L consecutive periods of no lost-sales of the $\underline{S}$-system. In our theoretical regret analysis, we gave a very strong *sufficient* condition, which is *independent* of the system state, to guarantee the occurrence of a triggering period (see Lemma 3 for details). Using this sufficient condition, we provided a *theoretical* upper bound on the time between two triggering periods by a geometric random variable of order $2L$. We emphasize again that this is a very loose theoretical upper bound and it works for *any* inventory state and for *any* period. However, in the actual simulation of these systems, we have observed that the triggering periods happen much more frequently, because these lost-sale events are not independent. For example, when the system has a very high on-hand inventory, then it is very likely that the system will experience no lost-sales for the next several periods. For the case of Gamma demand with mean 10 and shape 3, the empirical average lengths between triggering periods are merely $12, 32, 59, 90$ for $L = 5, 10, 15, 20$, respectively, that are much smaller than the theoretical upper bounds. We have tested examples with longer lead time L numerically, and observed similar patterns. However, we are unable to obtain an analytical expression for the average time between two triggering events because it depends on the joint distribution of the inventory state, which is extremely complex.

6. Concluding Remarks

In this paper, we proposed an improved nonparametric learning algorithm for the fundamental lost-sales inventory problem with positive lead times and censored demand, the simulated cycle-update (SCU) algorithm, and showed that its regret rate is $O(T^{1/2})$, which matches the lower bound of regret for any learning algorithms.

As its name suggests, the SCU algorithm constructs (random) cycles using a simulated system, and updates base-stock level at the beginning of each cycle. To overcome the challenges introduced by positive lead time and censored demand, we instituted two key ideas, namely, the withheld on-hand inventory and the double-phase gradient estimation. To analyze the performance of SCU algorithm, we introduced several bridging systems between the SCU-system and the optimal clairvoyant system. We also presented a simplified algorithm for the problem when the demand data is uncensored. Our numerical results demonstrated the effectiveness of the SCU algorithm.

We further comment on the benchmark used in this paper, which is the optimal base-stock policy. The major advantages of using this particular benchmark are (1) it is near-optimal when the ratio of the per-unit lost-sales penalty cost to the per-unit holding cost is large; (2) it makes the joint learning and optimization problem tractable by reducing the optimal policy space to be parametric with one parameter (base-stock level). The latter advantage, combined with the convexity result, allows us to design efficient and effective stochastic gradient descent type algorithms. For large lead times, one might consider using the best constant-order policy studied in Xin and Goldberg (2016) as an effective benchmark, as that has been shown to be asymptotically optimal.

We close this paper by pointing out several plausible directions for future research: (a) The model studied in this paper assumes that there is an infinite ordering capacity in each period. Many practical systems have finite ordering capacities. An interesting direction is to impose a finite ordering capacity constraint to the current problem, and develop a learning algorithm that

Distribution						$L=5$					$L=10$				
					T	100	200	1000	2000	5000	100	200	1000	2000	5000
Gamma	Shape	3	$p=$	50	SCU	31.1	29.8	13.4	8.0	3.7	18.5	24.9	21.0	15.2	8.3
					HJMR	3.5	3.9	4.1	3.5	2.7	18.4	23.3	22.8	18.0	11.1
					SCU-UN	25.4	21.3	7.8	4.5	2.0	19.4	23.1	14.8	9.6	4.7
				75	SCU	34.9	32.0	14.0	8.3	3.9	25.7	33.6	26.2	18.0	9.0
					HJMR	1.0	1.5	2.5	2.5	2.3	11.9	16.6	18.9	16.2	11.3
					SCU-UN	28.3	23.2	8.5	4.8	2.2	25.4	29.9	17.9	11.1	5.2
				100	SCU	35.9	33.2	14.3	8.6	4.3	30.6	38.8	28.8	19.0	9.3
					HJMR	0.2	0.7	1.7	1.9	1.8	7.7	11.8	15.6	13.9	10.8
					SCU-UN	29.4	24.2	9.1	5.3	2.5	28.3	32.7	18.8	11.5	5.3
		5		50	SCU	25.3	23.7	10.0	6.0	3.2	16.0	21.5	17.0	11.4	5.6
					HJMR	11.4	13.0	10.9	8.6	5.5	29.0	41.1	44.9	35.8	21.1
					SCU-UN	20.0	16.1	5.8	3.4	1.6	16.0	19.2	11.1	6.7	3.1
				75	SCU	27.1	25.2	11.1	6.8	4.3	21.7	27.8	20.1	12.9	6.2
					HJMR	7.3	8.8	9.0	7.4	5.2	19.8	30.3	41.8	37.6	26.8
					SCU-UN	21.1	17.8	6.8	4.1	2.0	20.1	23.6	12.9	7.7	3.5
				100	SCU	27.6	25.6	12.8	8.6	8.1	24.6	31.6	22.0	14.1	6.7
					HJMR	3.7	5.5	7.0	6.4	4.9	14.9	23.7	37.1	35.9	28.9
					SCU-UN	22.4	19.0	8.1	4.9	2.4	24.0	26.9	14.2	8.5	4.0
		7		50	SCU	21.7	20.5	9.0	5.7	4.2	14.1	19.1	14.2	9.2	4.4
					HJMR	18.5	23.3	20.2	15.1	8.8	36.0	53.1	65.5	54.4	32.4
					SCU-UN	16.5	14.0	5.4	3.2	1.6	14.3	17.1	9.1	5.4	2.5
				75	SCU	37.9	36.0	17.3	10.4	4.7	18.8	23.9	16.3	10.3	5.1
					HJMR	12.4	16.8	18.5	15.5	10.3	25.8	40.2	61.5	58.2	44.2
					SCU-UN	19.7	16.4	7.2	4.3	2.2	17.7	20.5	11.1	6.7	3.2
				100	SCU	38.4	38.9	18.4	11.3	5.3	20.8	26.3	18.1	11.7	6.2
					HJMR	9.3	13.1	17.0	15.1	10.9	19.6	31.9	56.4	57.0	49.1
					SCU-UN	19.5	18.4	10.0	6.1	3.1	20.2	22.9	12.7	8.0	3.9
Uniform				50	SCU	40.7	37.4	15.0	8.6	4.1	22.5	30.8	26.1	18.4	9.6
					HJMR	4.2	4.8	4.5	3.8	2.8	19.6	25.7	23.8	17.8	10.3
					SCU_UN	32.3	25.9	8.8	4.9	2.2	22.5	27.6	17.7	11.2	5.3
				75	SCU	41.8	37.4	14.9	8.8	5.1	30.4	39.5	30.7	20.2	9.8
					HJMR	1.9	2.4	3.2	3.0	2.5	13.2	18.5	21.4	18.0	12.1
					SCU_UN	33.8	26.9	9.3	5.2	2.4	29.2	34.9	19.7	11.8	5.3
				100	SCU	41.6	37.3	15.7	9.6	7.1	35.4	44.9	31.5	20.0	9.4
					HJMR	0.9	1.5	2.6	2.7	2.4	10.0	14.8	19.2	17.2	12.7
					SCU_UN	33.7	27.4	10.3	6.0	2.8	33.7	39.0	20.5	12.2	5.5
Poisson				50	SCU	19.4	19.7	10.7	9.7	6.3	13.1	18.4	13.3	8.4	4.7
					HJMR	29.5	39.1	38.0	28.7	15.7	44.0	66.7	93.5	82.1	51.5
					SCU_UN	16.1	14.2	6.7	4.1	2.1	13.0	16.3	8.8	5.4	2.6
				75	SCU	19.5	20.7	15.5	13.2	11.4	17.3	22.8	17.2	11.8	8.6
					HJMR	20.6	29.8	38.4	33.6	22.7	30.8	50.4	87.9	88.2	72.3
					SCU_UN	18.0	17.6	10.7	7.1	3.6	15.7	19.2	12.0	7.8	3.9
				100	SCU	20.3	23.0	22.7	14.5	13.0	18.9	25.4	21.0	15.3	13.9
					HJMR	16.0	24.6	36.6	34.7	26.7	23.9	40.8	81.1	86.3	79.7
					SCU_UN	18.3	19.0	13.1	8.9	4.5	18.5	21.7	15.4	10.8	5.7

Table 1 Performances (κ in %) of SCU, HJMR and SCU-UN when $L=5$ and $L=10$

converges to the clairvoyant's optimal modified base-stock policy. (b) The product considered in this paper is non-perishable. Many real-world products have a finite lifetime, and the resulting inventory system is known as the perishable inventory system. A possible research direction is to combine the ideas behind the SCU algorithm and the learning algorithm proposed in Zhang et al. (2018), and design an algorithm that can converge to the best base-stock policy for the perishable

Distribution						L=15					L=20				
					T	100	200	1000	2000	5000	100	200	1000	2000	5000
Gamma	Shape	3	$p=$	50	SCU	9.1	15.2	19.1	16.2	10.7	3.9	8.2	13.7	13.2	10.2
					HJMR	29.4	41.6	48.0	39.6	24.9	34.5	54.3	74.9	65.5	43.1
					SCU-UN	9.0	14.4	15.5	12.0	6.9	3.6	7.6	12.3	10.8	7.4
				75	SCU	13.7	22.6	28.2	23.0	14.0	7.5	14.5	24.3	22.2	15.9
					HJMR	19.5	29.8	42.8	39.4	29.1	23.3	38.6	64.9	63.6	50.7
					SCU-UN	14.0	22.2	22.0	15.8	8.4	7.2	14.2	20.6	17.2	10.5
				100	SCU	18.3	29.5	34.6	26.6	15.2	10.2	19.9	32.2	28.3	18.7
					HJMR	14.4	23.0	36.9	36.4	29.7	17.6	30.8	57.9	59.9	52.6
					SCU-UN	18.0	28.0	25.5	17.8	9.1	9.8	19.1	26.2	20.8	12.0
		5		50	SCU	7.2	12.7	16.1	13.0	7.7	3.5	7.2	12.9	11.5	7.7
					HJMR	37.8	58.4	79.8	69.4	44.3	40.1	69.3	82.7	74.9	52.0
					SCU-UN	7.7	12.4	12.3	8.8	4.5	3.3	7.3	10.8	8.7	5.1
				75	SCU	11.5	19.4	23.0	17.4	9.4	6.6	13.2	21.1	17.9	11.1
					HJMR	26.4	43.3	72.8	71.2	56.7	28.9	51.0	99.0	82.2	78.9
					SCU-UN	12.2	19.5	16.7	11.1	5.4	6.0	12.5	17.4	13.1	7.0
				100	SCU	14.9	24.6	26.5	19.2	10.0	8.5	16.6	26.6	22.1	13.0
					HJMR	20.0	34.3	65.5	68.1	60.5	22.2	41.4	89.1	65.6	55.4
					SCU-UN	14.5	22.8	19.1	12.3	5.9	8.6	16.5	21.4	15.4	7.9
		7		50	SCU	6.9	12.1	14.3	10.7	5.8	3.2	6.7	12.0	10.0	6.2
					HJMR	41.7	69.7	94.1	85.1	63.2	36.8	71.2	79.0	56.4	47.7
					SCU-UN	7.3	12.0	10.7	7.1	3.4	2.7	6.6	9.7	7.4	3.9
				75	SCU	10.3	17.9	19.5	13.8	7.0	6.1	12.2	18.9	15.2	8.7
					HJMR	31.4	52.9	97.5	99.0	83.2	32.2	58.6	85.0	64.1	50.7
					SCU-UN	10.3	16.8	14.3	9.1	4.4	5.6	11.2	15.5	11.0	5.6
				100	SCU	12.7	21.8	23.4	16.3	8.4	7.4	14.9	23.7	18.7	10.4
					HJMR	23.2	41.3	85.8	93.7	88.2	23.9	46.3	71.9	67.3	44.3
					SCU-UN	12.9	20.9	16.7	10.9	5.4	7.3	15.3	18.9	13.3	6.8
Uniform				50	SCU	11.5	18.8	23.7	19.9	13.0	4.4	9.7	17.8	17.1	12.9
					HJMR	29.4	42.8	49.7	39.9	23.5	34.9	55.5	76.1	66.2	41.8
					SCU_UN	10.8	17.7	19.4	14.7	8.2	4.6	10.6	15.9	13.7	9.1
				75	SCU	16.8	27.9	33.9	26.9	15.7	8.9	17.1	29.2	26.3	18.1
					HJMR	20.1	31.7	45.5	42.0	30.5	24.7	41.0	69.2	67.4	53.2
					SCU_UN	17.1	27.4	25.6	17.8	9.1	8.3	17.2	25.0	20.1	12.0
				100	SCU	20.6	33.9	39.8	29.8	16.2	11.3	22.2	37.7	32.7	21.1
					HJMR	15.4	25.1	42.0	41.7	33.9	18.1	32.2	61.5	64.4	57.1
					SCU_UN	19.4	31.2	28.4	19.3	9.4	11.2	22.4	30.6	23.6	13.2
Poisson				50	SCU	6.1	11.3	13.8	9.9	5.1	2.9	6.6	11.6	9.6	5.5
					HJMR	49.4	62.2	77.2	50.1	41.3	50.4	89.5	74.9	65.3	43.7
					SCU_UN	6.1	11.1	10.5	6.7	3.2	2.8	6.6	9.9	7.2	3.7
				75	SCU	9.7	16.7	19.4	13.9	7.6	4.7	10.7	19.2	15.9	9.3
					HJMR	34.2	59.8	88.3	71.8	58.6	35.1	65.6	75.7	60.1	42.0
					SCU_UN	9.6	16.0	14.2	9.9	5.1	5.0	11.0	15.8	11.4	5.9
				100	SCU	11.7	19.8	23.2	17.2	10.7	6.7	14.0	23.7	19.9	12.1
					HJMR	26.6	50.3	70.7	54.7	45.5	26.5	52.3	85.9	61.3	46.9
					SCU_UN	11.7	19.3	18.2	13.4	7.6	6.7	14.3	20.0	15.2	8.7

Table 2 Performances (κ in %) of SCU, HJMR and SCU-UN when $L = 15$ and $L = 20$

inventory system with positive lead time. Note that as a first step, some form of convexity result needs to be developed for that problem with respect to the base-stock levels, in order to adapt the online gradient descent approach. (c) Extend and expand the idea of (random) cycle-updating rule to other fundamental stochastic inventory problems where an ordering decision has a lasting effect on the underlying system, such as problems with setup costs, with ordering capacities, with random yield, or with non-linear purchasing costs.

Acknowledgments

The authors thank the department editor, the anonymous associate editor, and the anonymous referees for their very constructive and detailed comments, which have helped significantly improve both the content and the exposition of this paper. The research of Huanan Zhang is partially supported by NSF grants CMMI-1362619, CMMI-1634505 and CMMI-1634676. The research of Xiuli Chao is partially supported by NSF grants CMMI-1362619 and CMMI- 1634676. The research of Cong Shi is partially supported by NSF grants CMMI-1362619 and CMMI-1634505.

References

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

Bijvank, M., W.T. Huh, G. Jannakiraman, W. Kang. 2014. Robustness of order-up-to policies in lost-sales inventory systems. *Operational Research* **62**(5) 1049 – 1047.

Bijvank, M., I. F .A. Vis. 2011. Lost-sales inventory theory: A review. *European Journal of Operational Research* **215**(1) 1 – 13.

Burnetas, A. N., C. E. Smith. 2000. Adaptive ordering and pricing for perishable products. *Operations Research* **48**(3) 436–443.

Chen, W., M. Dawande, G. Janakiraman. 2014. Fixed-dimensional stochastic dynamic programs: An approximation scheme and an inventory application. *Operations Research* **62**(1) 81–103.

Goldberg, D. A., D. A. Katz-Rogozhnikov, Y. Lu, M. Sharma, M. S. Squillante. 2016. Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *Mathematics of Operations Research* **41**(3) 898–913.

Hazan, E. 2016. *Introduction to Online Convex Optimization*. Foundations and Trends in Optimization Series, Now Publishers, Boston, MA. URL http://ocobook.cs.princeton.edu/OCObook.pdf.

Huh, W. H., P. Rusmevichientong. 2009. A non-parametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research* **34**(1) 103–123.

Huh, W. T., G. Janakiraman, J. A. Muckstadt, P. Rusmevichientong. 2009a. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* **34**(2) 397–416.

Huh, W. T., G. Janakiraman, J. A. Muckstadt, P. Rusmevichientong. 2009b. Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science* **55**(3) 404–420.

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

Janakiraman, G., S. Seshadri, J. G. Shanthikumar. 2007. A comparison of the optimal costs of two canonical inventory systems. *Operations Research* **55**(5) 866–875.

Karlin, S., H. Scarf. 1958. *“Optimal Inventory Policy for the Arrow-Harris-Marschak Dynamic Model*. Stanford University Press, Stanford, California. In K. Arrow, S. Karlin, and H. Scarf (Eds.), Studies in the Mathematical Theory of Inventory and Production.

Kullback, S., R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1) 79–86.

Levi, R., G. Janakiraman, M. Nagarajan. 2008. A 2-approximation algorithm for stochastic inventory control models with lost-sales. *Mathematics of Operations Research* **33**(2) 351–374.

Lu, Y., M. S. Squillante, D. D. Yao. 2015. Matching supply and demand in production-inventory systems: Asymptotics and optimization. Working paper, IBM Thomas J. Watson Research Center, Yorktown Heights, NY.

Meyn, S.P., R.L. Tweedie. 1993. *Markov chains and stochastic stability*. Springer-Verlag, London.

Morton, T. E. 1969. Bounds on the solution of the lagged optimal inventory equation with no demand backlogging and proportional costs. *SIAM Review* **11**(4) 572–596.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization* **19**(4).

Philippou, A. N., C. Georghiou, G. N. Philippou. 1983. A generalized geometric distribution and some of its properties. *Statistics & Probability Letters* **1**(4) 171 – 175.

Reiman, M. I. 2004. A new and simple policy for the continuous review lost sales inventory model. *Working paper, Bell Laboratories, Murray Hill, NJ* .

Scheffe, H. 1947. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics* **18**(3) 434–438.

Shi, C., W. Chen, I. Duenyas. 2016. Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research* **64**(2) 362–370.

Tsybakov, A.B. 2009. *Introduction to Nonparametric Estimation*. Springer-Verlag, New York.

Xin, L., D. A. Goldberg. 2016. Optimality gap of constant-order policies decays exponentially in the lead time for lost sales models. *Operations Research* **64**(6) 1556–1565.

Zhang, H., X. Chao, C. Shi. 2018. Perishable inventory problems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research* **66**(5) 1276–1286.

Zipkin, P. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.

Zipkin, P. 2008a. Old and new methods for lost-sales inventory systems. *Operations Research* **56**(5) 1256–1263.

Zipkin, P. 2008b. On the structure of lost-sales inventory models. *Operations Research* **56**(4) 937–944.

Electronic Companion to Closing the Gap: A Learning Algorithm for Lost-Sales Inventory Systems with Lead Times

by Huanan Zhang, Xiuli Chao, and Cong Shi

Appendix A: An Example of the Evolution of the SCU- and the G-System

Example 1. In Figure 3, we use a simple example to illustrate how the dynamics of the SCU-system evolves and how it differs from the G-system. In this example, the lead time $L = 2$. We simulate the $\underline{S}$-system to determine the cycle length. Consider two cycles and suppose $S_2 > S_1$. In this case, the SCU policy will order more in the first period of cycle 2 to increase the inventory position from S_1 to S_2. Compared with the G-system, the inventory vectors of the two systems differ by 1 unit until τ'_2. In the fourth cycle, we have $S_4 < S_3$. In this case, the SCU policy marks 2 units of on-hand inventory as the withheld inventory. We can see that the withheld inventory amount keeps dropping, and apart from the withheld inventory, the two systems are the same. Note that in the second period of this cycle, by ignoring the withheld on-hand inventory, the SCU policy orders 4 (instead of 5), which is the same as what the G-system orders.

Appendix B: Proofs for Lemmas in §4

Proof of Lemma 1. It suffices to prove that for every sample path and in every period, after dropping the withheld on-hand inventory, every entry of the inventory vector of the SCU-system is no lower than that of the simulated $\underline{S}$-system.

From Theorem 1 in Huh et al. (2009a), we know that the inventory vector of a system operating under a base-stock level $S \geq \underline{S}$ is always no lower than that of the $\underline{S}$-system in all the entries. During the first cycle, since the SCU-system is the same as the base-stock system with $S_1 \geq \underline{S}$, the result clearly holds for the first cycle.

We prove the result for other periods using induction. Suppose the claim holds true from the first cycle to the $(k-1)$-th cycle for some $k \geq 2$, which is from period 1 to period $\tau_k - 1$. Then, we want to prove that the result is also true from period 1 to period $\tau_{k+1} - 1$.

Since the $\underline{S}$-system has no lost-sales from period $\tau_k - L$ to period $\tau_k - 1$, and the SCU-system has more on-hand inventory than the $\underline{S}$-system in these periods, then the pipeline inventory of the SCU-system at τ_k must be of the form $\left[\cdot, d_{\tau_k - 2}, \ldots, d_{\tau_k - L}\right]$, where the first entry (which is the order quantity in period τ_k) remains to be specified. There are two possible cases: 1) $I^{SCU}_{\tau_k} \geq I^{S_k}_{\tau_k} = S_k - \sum_{i=\tau_k - L}^{\tau_k - 1} d_i$ and 2) $I^{SCU}_{\tau_k} < I^{S_k}_{\tau_k}$.

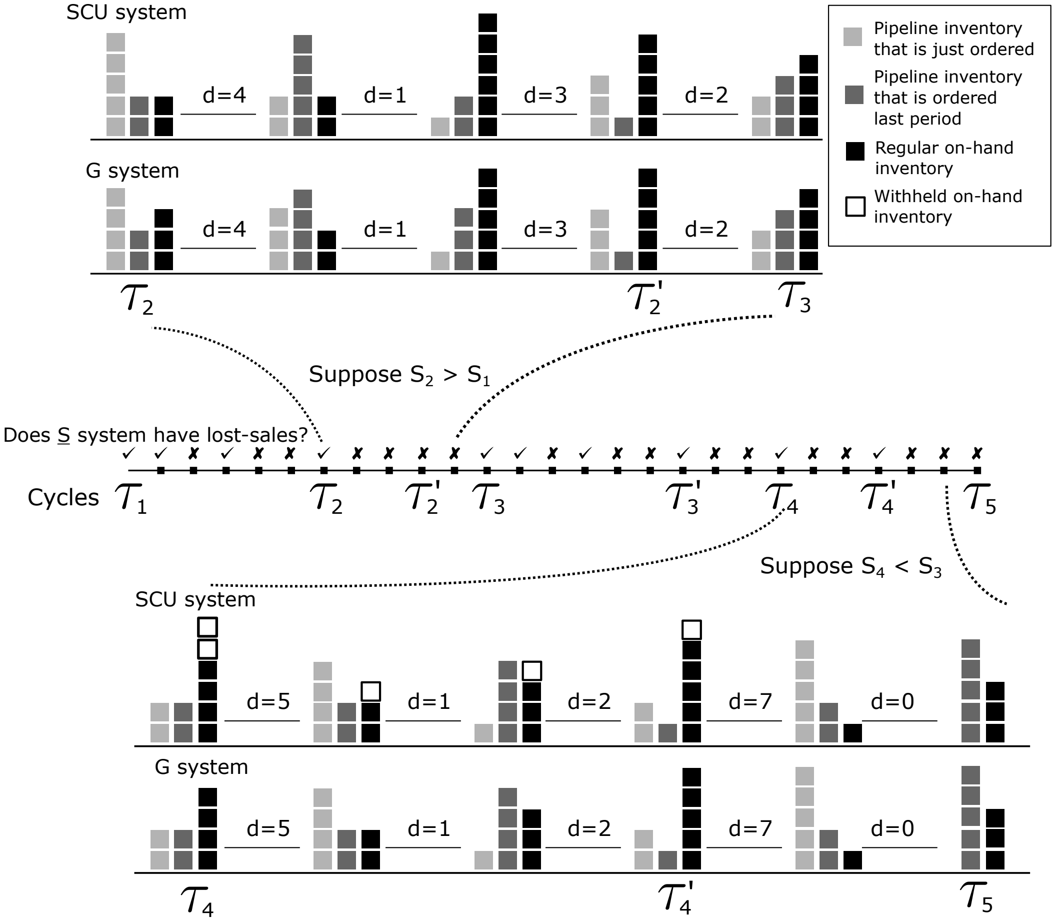

Figure 3 An graphical illustration of SCU policy with $L = 2$. For illustration purpose, all the numbers are integers

In Case 1), the SCU algorithm marks $I_{\tau_k} - S_k + \sum_{i=\tau_k-L}^{\tau_k-1} d_i$ amount of the on-hand inventory as withheld and orders d_{τ_k-1} in period τ_k. By the SCU algorithm, the regular (or non-withheld) inventory vector during this cycle in the SCU-system is the same as that of the base-stock system with the base-stock level S_k, which we shall refer to as the S_k-system. Now, comparing the S_k-system and the $\underline{S}$-system at the beginning of period τ_k, the only difference lies in their on-hand inventory levels, namely, $S_k - \sum_{i=\tau_k-L}^{\tau_k-1} d_\ell$ and $\underline{S} - \sum_{i=\tau_k-L}^{\tau_k-1} d_\ell$, which are achieved in period τ_k when both systems started out empty in period 1 and followed their own base-stock policies. By the monotonicity result in Theorem 1 of Huh et al. (2009a), this implies that the inventory vector for the S_k-system is no lower than that of the $\underline{S}$-system between period τ_k and period $\tau_{k+1} - 1$. Hence, the result follows from the fact that the inventory vector of the SCU-system is no lower than that of the S_k-system, and that the inventory vector of the S_k-system is no lower than that of the $\underline{S}$-system during the k-th cycle.

In Case 2), the SCU algorithm orders

$$d_{\tau_k - 1} + S_k - \sum_{i=\tau_k - L}^{\tau_k - 1} d_i - I_{\tau_k} = S_k - \sum_{i=\tau_k - L}^{\tau_k - 2} d_i - I_{\tau_k}$$

in period τ_k to bring the inventory position up to S_k. Now consider another system that starts in period τ_k with the same inventory vector as the SCU-system, but implements a base-stock policy S_{k-1} between period τ_k and period $\tau_{k+1} - 1$. With a slight abuse of notation, call the latter system the S_{k-1}-system. Then, it can be seen that the inventory vector of the SCU-system between period τ_k and period $\tau_{k+1} - 1$ is always no lower than that of the S_{k-1}-system. On the other hand, by applying Theorem 1 in Huh et al. (2009a) to the S_{k-1}-system and similar arguments as in Case 1) above, we can show that the inventory vector in the S_{k-1}-system is no lower than that of the $\underline{S}$-system between period τ_k and period $\tau_{k+1} - 1$. This proves that the inventory vector of the SCU-system is no lower than that of the $\underline{S}$-system during cycle k.

Thus, the result holds for both cases during the k-th cycle. This completes the induction argument and the proof of Lemma 1. **Q.E.D.**

Proof of Lemma 2. We consider the following two cases: 1) $I^{SCU}_{\tau_k} \geq I^{S_k}_{\tau_k}$, and 2) $I^{SCU}_{\tau_k} < I^{S_k}_{\tau_k}$, separately.

First suppose that $I^{SCU}_{\tau_k} \geq I^{S_k}_{\tau_k}$. In this case, it follows from $\hat{I}^{SCU}_{\tau_k} := \left(\hat{I}^{SCU}_{\tau_k} - (S_k - S_{k-1})\right)^+$ that there are two sources of excess withheld on-hand inventory at the beginning of cycle k in the SCU-system. The first is inherited from the previous cycle, and the second is the newly created ones due to a decrease in the target base-stock level. Since the pipeline inventories at the beginning of cycle k in both SCU- and G-systems are equal (as both experienced no lost-sales for L consecutive periods), the SCU-system and the G-system would be the same in period τ_k if the withheld on-hand inventory is ignored. Furthermore, since the withheld on-hand inventory is not counted when making ordering decisions in SCU-system, the ordering quantities of the two systems would be the same within this cycle, and the SCU system will always have no lower on-hand inventory than the G-system for each period in both the first and second phase of cycle k.

Next suppose that $I^{SCU}_{\tau_k} < I^{S_k}_{\tau_k}$. In this case, the sales data collected from SCU-system may not allow us to simulate the G-system as it has lower on-hand inventory in period τ_k (and maybe also in some subsequent periods). Since both SCU- and G-systems implement the base-stock level S_k, it follows from Lemma 1 above and Huh et al. (2009a) that both systems would have experienced no lost-sales for L consecutive periods before the triggering period τ'_k. This implies that, the inventory state of SCU- and G- systems, after placing order, will be both equal to $\left(d_{\tau_k - 1}, \ldots, d_{\tau'_k - L}, S_k - \sum_{t=\tau'_k - L}^{\tau'_k - 1} d_t\right)$. This show that the SCU- and G- systems would be identical during the second phase of cycle k, and in particular, they will have the same on-hand inventory level in each period of the second phase of cycle k.

Combining the two cases, we complete the proof of Lemma 2. **Q.E.D.**

Proof of Lemma 3. Denote the lost-sales from period a to period b in $\underline{S}$ system as $m^{\underline{S}}[a,b]$, then under the stated condition, we have, for any $k = t+L, \ldots, t+2L-1$,

$$I_k^{\underline{S}} = \underline{S} - d_{[k-L,k-1]} + m^{\underline{S}}[a,b] \geq \underline{S} - d_{[k-L,k-1]} \geq \frac{\underline{S}}{L+1} \geq d_k.$$

This implies that there will be no lost-sales from period $t+L$ to $t+2L-1$. **Q.E.D.**

For a base-stock system with base-stock level S, if $d_k > \frac{S+\delta}{L+1}$ for $k = t, \ldots, t+L$, then the total lost-sales amount from period t to period $t+L$ is at least δ.

Proof of Lemma 4. Because all the pipeline inventory in period t will be available in the beginning of period $t+L$, the starting inventory level in period $t+L$ will be $I_{t+L} = S - \sum_{k=t}^{t+L-1} d_k + \sum_{k=t}^{t+L-1}(d_k - I_k)^+$, where $\sum_{k=t}^{t+L-1}(d_k - I_k)^+$ is the total lost-sales amount from period t to $t+L-1$. The lost-sales amount in period $t+L$ be $(d_{t+L} - I_{t+L})^+$. Then we have

$$\sum_{k=t}^{t+L}(d_k - I_k)^+ = \sum_{k=t}^{t+L-1}(d_k - I_k)^+ + \left(d_{t+L} - S + \sum_{k=t}^{t+L-1} d_k - \sum_{k=t}^{t+L-1}(d_k - I_k)^+\right)^+ \geq \sum_{k=t}^{t+L} d_k - S \geq \Delta.$$

Q.E.D.

Proof of Lemma 5. Since

$$|I_t^o - I_t^\beta| = (I_t^o - I_t^\beta)^+ + (I_t^\beta - I_t^o)^+,$$

and at most one term on the right hand side can be positive, it suffices to prove, for all $t \geq 1$,

$$(I_t^o - I_t^\beta)^+ \leq \beta, \quad (I_t^\beta - I_t^o)^+ \leq \beta.$$

In the following, we prove, by induction, that much stronger results hold: For all $t \geq 1$,

$$(I_t^o - I_t^\beta)^+ + \sum_{i=0}^{L-1}(q_{t-i}^o - q_{t-i}^\beta)^+ \leq \beta, \tag{18}$$

$$(I_t^\beta - I_t^o)^+ + \sum_{i=0}^{L-1}(q_{t-i}^\beta - q_{t-i}^o)^+ \leq \beta. \tag{19}$$

By our definition of the original and β-system, (18) and (19) are clearly satisfied when $t=1$. Suppose (18) and (19) hold at t, we will show that (18) and (19) continue to hold at $t+1$.

We first focus on (18). By the system dynamics of base-stock policy S, we have

$$\begin{aligned}&(I^o_{t+1}-I^\beta_{t+1})^+ + \sum_{i=0}^{L-1}(q^o_{t+1-i}-q^\beta_{t+1-i})^+ \qquad (20)\\ &= \big((I^o_t-d_t)^+-(I^\beta_t-d_t)^+ + q^o_{t-L+1}-q^\beta_{t-L+1}\big)^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+ + \big(\min(I^o_t,d_t)-\min(I^\beta_t,d_t)\big)^+.\end{aligned}$$

We prove (18) holds by considering four cases separately: 1) $d_t\ge\max(I^o_t,I^\beta_t)$, 2) $d_t\le\min(I^o_t,I^\beta_t)$, 3) $I^o_t\le d_t\le I^\beta_t$, and 4) $I^\beta_t\le d_t\le I^o_t$.

Case 1): By (20), the left hand side of (18) at $t+1$ is

$$\begin{aligned}&\big((I^o_t-d_t)^+-(I^\beta_t-d_t)^+ + q^o_{t-L+1}-q^\beta_{t-L+1}\big)^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+ + \big[\min(I^o_t,d_t)-\min(I^\beta_t,d_t)\big]^+\\ &= (q^o_{t-L+1}-q^\beta_{t-L+1})^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+ + (I^o_t-I^\beta_t)^+\\ &= (I^o_t-I^\beta_t)^+ + \sum_{i=0}^{L-1}(q^o_{t-i}-q^\beta_{t-i})^+\\ &\le \beta,\end{aligned}$$

where the inequality follows from the inductive assumption.

Case 2): In this case, we have, by (20),

$$\begin{aligned}&\big((I^o_t-d_t)^+-(I^\beta_t-d_t)^+ + q^o_{t-L+1}-q^\beta_{t-L+1}\big)^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+ + \big(\min(I^o_t,d_t)-\min(I^\beta_t,d_t)\big)^+\\ &= \big(I^o_t-I^\beta_t + q^o_{t-L+1}-q^\beta_{t-L+1}\big)^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+\\ &\le (I^o_t-I^\beta_t)^+ + (q^o_{t-L+1}-q^\beta_{t-L+1})^+ + \sum_{i=0}^{L-2}(q^o_{t-i}-q^\beta_{t-i})^+\\ &= (I^o_t-I^\beta_t)^+ + \sum_{i=0}^{L-1}(q^o_{t-i}-q^\beta_{t-i})^+\\ &\le \beta,\end{aligned}$$

where the first inequality follows from $(a+b)^+\le a^+ + b^+$ for any real numbers a and b, and the second inequality follows from the inductive assumption.

Case 3): This case can happen only when $I_t^o \le I_t^\beta$. We have

$$\begin{aligned}
&\big((I_t^o - d_t)^+ - (I_t^\beta - d_t)^+ + q_{t-L+1}^o - q_{t-L+1}^\beta\big)^+ + \sum_{i=0}^{L-2}(q_{t-i}^o - q_{t-i}^\beta)^+ + \big(\min(I_t^o, d_t) - \min(I_t^\beta, d_t)\big)^+ \\
= &\big(q_{t-L+1}^o - q_{t-L+1}^\beta - (I_t^\beta - d_t)^+\big)^+ + \sum_{i=0}^{L-2}(q_{t-i}^o - q_{t-i}^\beta)^+ \\
\le &\sum_{i=0}^{L-1}(q_{t-i}^o - q_{t-i}^\beta)^+ \\
\le &\beta,
\end{aligned}$$

where the equality follows from $\min(I_t^o, d_t) - \min(I_t^\beta, d_t) \le 0$ (and hence the last term is 0), the first inequality follows from $(a-b)^+ \le a^+$ for any real numbers a and $b \ge 0$, and the second inequality follows from the inductive assumption.

Case 4): This last case can occur when $I_t^o \ge I_t^\beta$, and we have

$$\begin{aligned}
&\big((I_t^o - d_t)^+ - (I_t^\beta - d_t)^+ + q_{t-L+1}^o - q_{t-L+1}^\beta\big)^+ + \sum_{i=0}^{L-2}(q_{t-i}^o - q_{t-i}^\beta)^+ + \big(\min(I_t^o, d_t) - \min(I_t^\beta, d_t)\big)^+ \\
= &\big((I_t^o - d_t) + q_{t-L+1}^o - q_{t-L+1}^\beta\big)^+ + \sum_{i=0}^{L-2}(q_{t-i}^o - q_{t-i}^\beta)^+ + (d_t - I_t^\beta) \\
\le &(I_t^o - d_t) + (q_{t-L+1}^o - q_{t-L+1}^\beta)^+ + \sum_{i=0}^{L-2}(q_{t-i}^o - q_{t-i}^\beta)^+ + (d_t - I_t^\beta) \\
= &(I_t^o - I_t^\beta)^+ + \sum_{i=0}^{L-1}(q_{t-i}^o - q_{t-i}^\beta)^+ \\
\le &\beta,
\end{aligned}$$

where again we used $(a+b)^+ \le a^+ + b^+$ in the first inequality, and the second inequality is by the inductive assumption.

Hence (18) is satisfied for $t+1$ as well. Similar argument proves (19) for $t+1$. This finishes the induction proof and the proof of Lemma 5. **Q.E.D.**

Proof of Lemma 6. Recall that τ_k, τ_k' and τ_{k+1} are three adjacent triggering periods defined by the $\underline{S}$-system. Thus, $\nabla G(S_k, \tau_k, \tau_k' - 1)$ and $\nabla G(S_k, \tau_k', \tau_{k+1} - 1)$ are determined by $[d_{\tau_k-1}, d_{\tau_k-2}, \ldots, d_{\tau_k-L}]$ and $[d_{\tau_k'-1}, d_{\tau_k'-2}, \ldots, d_{\tau_k'-L}]$, respectively. If we re-index $\{\tau_1, \tau_2, \tau_2', \tau_3, \tau_3', \ldots\}$ as $\{r(1), r(2), r(3), r(4), r(5), \ldots\}$, then the process $\{\mathbf{d}_i = [d_{r(i)-1}, d_{r(i)-2}, \ldots, d_{r(i)-L}]; i \ge 1\}$ is a Markov chain on a general state space (or a Harris chain). It is important to keep in mind that this Markov chain is solely determined by the $\underline{S}$-system, and it is not affected by the SCU- or the G-system.

We show that, under Assumption 1, $\{\mathbf{d}_i; i \geq 1\}$ is ergodic and converges to a stationary distribution $\mathbf{d}_\infty$ exponentially fast. Following the approach in Huh et al. (2009a), we use uniform ergodicity to prove this result. A measurable set $\mathbf{U} \subseteq \mathbb{R}_+^L$ is called a *small set* with respect to a nontrivial measure ν, if there exists an $i^* > 0$ such that for any $\mathbf{d} \in \mathbf{U}$ and any measurable set $B = (B_1, \ldots, B_L) \subseteq \mathbb{R}_+^L$, it holds that $\mathbb{P}(\mathbf{d}_{i^*} \in B \mid \mathbf{d}_1 = \mathbf{d}) \geq \nu(B)$. By Theorem 16.0.2 of Meyn and Tweedie (1993), if $\mathbf{U}$ is a small set with respect to ν, then there exists a stationary random variable $\mathbf{d}_\infty$ such that for any $\mathbf{d} \in \mathbf{U}$ and $i \geq i^*$, it satisfies $\delta_{i+1}(\mathbf{d}) \leq (1 - \nu(\mathbb{R}_+^L))^{\frac{i}{i^*-1}}$, where

$$\delta_i(\mathbf{d}) = \sup_B \left\{ |\mathbb{P}(\mathbf{d}_i \in B \mid \mathbf{d}_1 = \mathbf{d}) - \mathbb{P}(\mathbf{d}_\infty \in B)| : \text{measurable set } B \subseteq \mathbb{R}_+^L \right\}.$$

By the Scheffe's Theorem (Scheffe (1947)), we have

$$\delta_i(\mathbf{d}) = \frac{1}{2} \int_{\mathbf{z}} \left| \big(P(\mathbf{d}_i \in d\mathbf{z} | \mathbf{d}_1 = \mathbf{d}) - P(\mathbf{d}_\infty \in d\mathbf{z}) \big) \right|. \tag{21}$$

The first step is to define $\mathbf{U}$, B, ν and i^* for our Markov chain. Since $\mathbf{d}$ is the pipeline inventory of the $\underline{S}$-system at the beginning of a triggering period, we must have $\mathbf{d} \cdot \mathbf{1}^L \leq \underline{S}$, where $\mathbf{d} \cdot \mathbf{1}^L$ is the sum of all entries of $\mathbf{d}$. Let $\mathbf{U} = \{\mathbf{d} \in \mathbb{R}_+^L \mid \mathbf{d} \cdot \mathbf{1}^L \leq \underline{S}\}$, and B_k be any measurable set in $\mathbb{R}_+$ for $k = 1, \ldots, L$, and denote $B = B_1 \times \cdots \times B_L$. Define

$$\nu(B) = \left(\mathbb{P}\left(D \in \left(\cap_{k=1}^L B_k \right) \cap \left[0, \frac{\underline{S}}{L+1} \right] \right) \right)^{2L},$$

where D represents a generic demand. We now prove that $\mathbf{U}$ is a small set with respect to ν and $i^* = 2$, i.e., for any $\mathbf{d} \in \mathbf{U}$ and $B \in \mathbb{R}_+^L$, we have $\mathbb{P}\big(\mathbf{d}_2 \in B \mid \mathbf{d}_1 = \mathbf{d}\big) \geq \nu(B)$.

Consider the event that the demands in periods $1, 2, \ldots, 2L$ satisfy

$$E = \left\{ D_k \in \left(\cap_{k=1}^L B_k \right) \cap \left[0, \frac{\underline{S}}{L+1} \right], \text{ for } k = 1, 2, \ldots, 2L \right\}.$$

By Lemma 3, for any initial state $\mathbf{d}_1 = \mathbf{d}$, there is no lost sales in the $\underline{S}$-system from periods L to $2L$, implying $r(2) \leq 2L$. Moreover, on the event E, it is seen that the pipeline inventory at the beginning of period $r(2)$ satisfies $\mathbf{d}_2 \in B$. Hence, for any $\mathbf{d} \in \mathbf{U}$, we have

$$\mathbb{P}(\mathbf{d}_2 \in B \mid \mathbf{d}_1 = \mathbf{d}) \geq \mathbb{P}(E) = \nu(B).$$

This shows that the Markov chain $\{\mathbf{d}_k;\ k \geq 1\}$ is uniformly ergodic. Applying Theorem 16.0.2 of Meyn and Tweedie (1993), we obtain, for all $i \geq 2$,

$$\delta_i(\mathbf{d}) \leq (1 - c_1^{2L})^{i-1}. \tag{22}$$

For notational convenience, we define

$$H(d_{\tau_k-1}, d_{\tau_k-2}, \dots, d_{\tau_k-L}) = \mathbb{E}[\nabla G(S_k, \tau_k, \tau_k'-1) \big| d_{\tau_k-1}, d_{\tau_k-2}, \dots, d_{\tau_k-L}].$$

Then for any $[d_{\tau_k-1}, d_{\tau_k-2}, \dots, d_{\tau_k-L}]$, $H(d_{\tau_k-1}, d_{\tau_k-2}, \dots, d_{\tau_k-L})$ is upper bounded by

$$H(d_{\tau_k-1}, d_{\tau_k-2}, \dots, d_{\tau_k-L}) \le \frac{1-c_1^L}{(1-c_1)c_1^L}\max(h,b). \tag{23}$$

Therefore, we have

$$\begin{aligned}
\big|\mathbb{E}[H(\mathbf{d}_k) - \mathbb{E}[H(\mathbf{d}_\infty)]\big| &= \Big|\int_{\mathbf{z}} H(\mathbf{z})\big(P(\mathbf{d}_i \in d\mathbf{z}|\mathbf{d}_1 = \mathbf{d}) - P(\mathbf{d}_\infty \in d\mathbf{z})\big)\Big| \\
&\le \int_{\mathbf{z}} H(\mathbf{z})\big|\big(P(\mathbf{d}_i \in d\mathbf{z}|\mathbf{d}_1 = \mathbf{d}) - P(\mathbf{d}_\infty \in d\mathbf{z})\big)\big| \\
&\le \frac{1-c_1^L}{(1-c_1)c_1^L}\max(h,b)\int_{\mathbf{z}}\big|\big(P(\mathbf{d}_i \in d\mathbf{z}|\mathbf{d}_1 = \mathbf{d}) - P(\mathbf{d}_\infty \in d\mathbf{z})\big)\big| \\
&= \frac{1-c_1^L}{(1-c_1)c_1^L}\max(h,b)\cdot 2\delta_k(\mathbf{d}) \\
&\le \frac{1-c_1^L}{(1-c_1)c_1^L}\max(h,b)\cdot 2(1-c_1^{2L})^{k-1} \\
&= o\left(\frac{1}{\sqrt{k}}\right),
\end{aligned}$$

where the second inequality follows (23), and the second equality follows from (21), and the last inequality follows from (22), and the last equality follows from the fact that ρ^k tends to 0 faster than $1/\sqrt{k}$ for any $\rho \in (0,1)$.

Applying the above result, we obtain

$$\begin{aligned}
&\Big|\mathbb{E}[\nabla G(S_k, \tau_k, \tau_k'-1)] - \mathbb{E}[\nabla G(S_k, \tau_k', \tau_{k+1}-1)]\Big| \\
=& \Big|\mathbb{E}\big[\mathbb{E}[\nabla G(S_k, \tau_k, \tau_k'-1)|\mathbf{d}_k]\big] - \mathbb{E}\big[\mathbb{E}[\nabla G(S_k, \tau_k', \tau_{k+1}-1)|\mathbf{d}_{k+1}]\big]\Big| \\
=& \big|\mathbb{E}[H(\mathbf{d}_k)] - \mathbb{E}[H(\mathbf{d}_{k+1})]\big| \\
=& \Big|\big(\mathbb{E}[H(\mathbf{d}_k)] - \mathbb{E}[H(\mathbf{d}_\infty)]\big) - \big(\mathbb{E}[H(\mathbf{d}_{k+1})] - \mathbb{E}[H(\mathbf{d}_\infty)]\big)\Big| \\
=& o(1/\sqrt{k}).
\end{aligned}$$

This completes the proof of (10). **Q.E.D.**

Proof of Lemma 7. We have

$$\mathbb{E}\left[\sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1}-1)\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{S^*}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1}-1)\right] - \mathbb{E}\left[\sum_{k=1}^{N} G(S^*, \tau_k, \tau_{k+1}-1)\right] \\
&\geq \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}\left(S_k > S^* + 1/2\right)\left[G(S_k, \tau_k, \tau_{k+1}-1) - G(S^*, \tau_k, \tau_{k+1}-1)\right]\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}\left(S_k > S^* + 1/2\right)\left[G(S_k, \tau_\infty, \tau_{\infty+1}-1) - G(S^*, \tau_\infty, \tau_{\infty+1}-1)\right]\right] + o(\sqrt{T}) \\
&\geq \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}\left(S_k > S^* + 1/2\right)\left[G(S^* + 1/2, \tau_\infty, \tau_{\infty+1}-1) - G(S^*, \tau_\infty, \tau_{\infty+1}-1)\right]\right] + o(\sqrt{T}) \\
&= \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}\left(S_k > S^* + 1/2\right)\right] \cdot \mathbb{E}\left[G(S^* + 1/2, \tau_\infty, \tau_{\infty+1}-1) - G(S^*, \tau_\infty, \tau_{\infty+1}-1)\right] + o(\sqrt{T}).
\end{aligned}
$$

The second equality follows from the argument of Lemma 6, by applying the convergence result to the function G instead of ∇G. The second inequality is due to convexity of $\mathbb{E}\left[G(\cdot, \tau_\infty, \tau_{\infty+1}-1)\right]$ so that $G(S_k, \tau_\infty, \tau_{\infty+1}-1) \geq G(S^* + 1/2, \tau_\infty, \tau_{\infty+1}-1)$ for any $S_k > S^* + 1/2$.

We can then readily prove the result by contradiction. Now suppose that

$$
\mathbb{E}\left[\sum_{k=1}^{N} \mathbb{1}\left(S_k > S^* + 1/2\right)\right] > O(\sqrt{T}),
$$

which immediately implies that

$$
\mathbb{E}\left[\sum_{k=1}^{N} G(S_k, \tau_k, \tau_{k+1}-1)\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{S^*}\right] > O(\sqrt{T}),
$$

which contradicts Proposition 2. This completes the proof. **Q.E.D.**

Appendix C: Proof for the Lower Bound (Proposition 1)

For the discrete demand case, the discrete demand example provided in Besbes and Muharremoglu (2013) gives the desired lower bound. For the continuous case, we provide an example with continuous demand and show that its regret under any learning policy is lower bounded by $\Omega(\sqrt{T})$. This example is a slight modification of the discrete demand example provided in Besbes and Muharremoglu (2013), and the lower bound proof also follows their argument.

Example 2. Consider an inventory control problem with lost sales and $h = p = 1$, $L = 0$ and $T > 5$. The demand follows one of two potential distributions, with the cdf F_a and F_b given by

$$F_a(x) = \begin{cases} (\frac{1}{8} + \frac{1}{4\sqrt{T}})x & \text{for } 0 \le x < 4, \\ \frac{1}{2} + \frac{1}{\sqrt{T}} & \text{for } 4 \le x < 400, \\ (\frac{1}{8} - \frac{1}{4\sqrt{T}})(x-400) + \frac{1}{2} + \frac{1}{\sqrt{T}} & \text{for } 400 \le x < 404, \\ 1 & \text{for } x \ge 404, \end{cases}$$

and

$$F_b(x) = \begin{cases} (\frac{1}{8} - \frac{1}{4\sqrt{T}})x & \text{for } 0 \le x < 4, \\ \frac{1}{2} - \frac{1}{\sqrt{T}} & \text{for } 4 \le x < 400, \\ (\frac{1}{8} + \frac{1}{4\sqrt{T}})(x-400) + \frac{1}{2} - \frac{1}{\sqrt{T}} & \text{for } 400 \le x < 404, \\ 1 & \text{for } x \ge 404. \end{cases}$$

Then, the optimal base-stock level for F_a, denoted by S_a^*, is within $(2,4)$, and the optimal base-stock level for F_b, denoted by S_b^*, is within $(400,402)$. Note that this satisfies our Assumption 1, by setting $\underline{D} = \underline{S} = 2$ and $\bar{D} = \bar{S} = 402$ and $c_1 = 0.13$ and $c_2 = 0.06$.

We prove that, even with observable demand, no policy can achieve a worst-case expected regret better than $\Omega(\sqrt{T})$.

Let π be an arbitrary policy. The worst-case expected regret of π is bounded from below by

$$(h+p)\frac{196}{\sqrt{T}} \max\left\{\sum_{t=1}^{T} \mathbb{P}_a^\pi\left(S_t^\pi(\omega) > 200\right),\ \sum_{t=1}^{T} \mathbb{P}_b^\pi\left(S_t^\pi(\omega) \le 200\right)\right\}, \tag{24}$$

where $S_t^\pi(\omega)$ is the base-stock level in period t for policy π under sample path ω. The term $\max\left\{\sum_{t=1}^{T} \mathbb{P}_a^\pi\left(S_t^\pi(\omega) > 200\right),\ \sum_{t=1}^{T} \mathbb{P}_b^\pi\left(S_t^\pi(\omega) \le 200\right)\right\}$ in (24) provides a lower bound of periods that the base-stock level used by policy π is at least 196 away from the optimal base-stock level. When the underlying demand is $F_a(x)$, increasing the base-stock level from 4 to 200 will at least increase the cost by $(h+p)\frac{196}{\sqrt{T}}$. Similarly, when the underlying demand is $F_b(x)$, decreasing the base-stock level from 400 to 200 will at least increase the cost by $(h+p)\frac{200}{\sqrt{T}}$. Therefore, we obtain (24) as a lower bound of the regret, which can be further lower bounded by

$$(h+p)\frac{196}{2\sqrt{T}} \sum_{t=1}^{T} \max\left\{\mathbb{P}_a^\pi\left(S_t^\pi(\omega) > 200\right),\ \mathbb{P}_b^\pi\left(S_t^\pi(\omega) \le 200\right)\right\}. \tag{25}$$

By Theorem 2.2 in Tsybakov (2009), we have

$$\max\left\{\mathbb{P}_a^\pi\left(S_t^\pi(\omega) > 200\right), \mathbb{P}_b^\pi\left(S_t^\pi(\omega) \le 200\right)\right\} \ge \frac{1}{4} \cdot \exp\{-\mathcal{K}_{t-1}(\mathbb{P}_a, \mathbb{P}_b)\}, \tag{26}$$

where

$$\mathcal{K}_t(\mathbb{P}_a, \mathbb{P}_b) = \mathbb{E}_a\left[\log \frac{\mathbb{P}_a(D_1, \ldots, D_t)}{\mathbb{P}_b(D_1, \ldots, D_t)}\right]$$

is the Kullback-Leibler divergence (see Kullback and Leibler (1951)) between the distributions of $\{D_1, \ldots, D_t\}$ under F_a and under F_b, which is equal to

$$\mathcal{K}_t(\mathbb{P}_a, \mathbb{P}_b) = t\left[\left(\frac{1}{2} + \frac{1}{\sqrt{T}}\right)\log\left(\frac{1+\frac{2}{\sqrt{T}}}{1-\frac{2}{\sqrt{T}}}\right) + \left(\frac{1}{2} - \frac{1}{\sqrt{T}}\right)\log\left(\frac{1-\frac{2}{\sqrt{T}}}{1+\frac{2}{\sqrt{T}}}\right)\right]. \qquad (27)$$

It is a simple exercise to show that $2x \leq \log\frac{1+x}{1-x} \leq 2x + 2x^2$ for $x \in (0, 1/2)$. Substituting this inequality to (27) above, we obtain $\mathcal{K}_t(\mathbb{P}_a, \mathbb{P}_b) \leq \frac{7t}{T}$. Plugging this into (26) yields

$$\max\left\{\mathbb{P}_a^\pi\left(S_t^\pi(\omega) > 200\right),\ \mathbb{P}_b^\pi\left(S_t^\pi(\omega) \leq 200\right)\right\} \geq \frac{1}{4}\exp\left\{-\frac{7(t-1)}{T}\right\} \geq \frac{1}{4}e^{-7}.$$

Consequently, (25) is bounded from below by

$$(h+p)\,\frac{196}{2\sqrt{T}}\sum_{t=1}^{T}\frac{1}{4}e^{-7} = 24e^{-7}\sqrt{T}.$$

This completes the proof of Proposition 1. **Q.E.D.**