Perishable Inventory Systems: Convexity Results for Base-Stock Policies and Learning Algorithms under Censored Demand

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 develop the first nonparametric learning algorithm for periodic-review perishable inventory systems. In contrast to the classical perishable inventory literature, we assume that the firm does not know the demand distribution *a priori* and makes replenishment decision in each period based only on the past sales (censored demand) data. It is well-known that even with complete information about the demand distribution *a priori*, the optimal policy for this problem does not possess a simple structure. Motivated by the studies in the literature showing that base-stock policies perform near-optimal in these systems, we focus on finding the best base-stock policy. We first establish a convexity result, showing that the total holding, lost-sales and outdating cost is convex in the base-stock level. Then, we develop a nonparametric learning algorithm that generates a sequence of order-up-to levels whose running average cost converges to the cost of the optimal base-stock policy. We establish a square-root convergence rate of the proposed algorithm, which is the best possible. Our algorithm and analyses require a novel method for computing a valid cycle subgradient and the construction of a bridging problem, which significantly departs from previous studies.

Key words: inventory, perishable products, base-stock policy, censored demand, learning algorithms

Received March 2016; revisions received July 2017, November 2017; accepted December 2017 by Operations Research.

1. Introduction

Perishable products are undoubtedly an indispensable part of our lives. For example, perishable products such as meat, fruit, vegetable, dairy products, and frozen foods constitute the majority of supermarket sales. Moreover, virtually all pharmaceuticals belong to the category of perishable products. Perhaps the most frequently discussed applications of perishable inventory models are inventory control of blood products in blood banks (see Prastacos (1984)). To this date, almost all the papers on this topic assume that the stochastic demand processes are given as an input to the models, and the inventory replenishment decisions are made with full knowledge of the demand distribution. However, in practice, the underlying demand distribution may not be available to the firm *a priori*. This raises a natural and important research question as to how to devise an efficient and effective learning algorithm that only uses the observed sales data collected over time, so as to minimize the total expected cost.

1.1. Model, Benchmark, and Objective

Consider a periodic-review stochastic inventory systems with perishable products. The product lifetime is m periods. The demands across periods are i.i.d. continuous random variables. The firm makes a replenishment decision at the beginning of each period, and then demand is realized and satisfied to the maximum extent from the on-hand inventory. We consider the class of First-In-First-Out (FIFO) issuing policies, i.e., the oldest inventory is consumed first when demand arrives (see Karaesmen et al. (2011)). Any unsatisfied demand is *lost*, any product reaching the end of its lifetime expires, while non-expired excess inventory is carried over to the next period. Besides the typical inventory holding cost and lost sales penalty cost, there is an outdating cost for each inventory unit that is wasted due to expiration.

However, contrary to the classical perishable inventory literature, the underlying demand distribution D_t is not known to the firm *a priori*. Instead, the firm makes ordering decisions based on observed past sales, which is the minimum of the realized demand and the on-hand inventory. In other words, the sales data are *censored* by the available inventory level, and the firm cannot observe the lost-sales quantity. We note that joint learning and optimization problems under censored demand information for non-perishable inventory systems have received much attention in the research literature and are often challenging to analyze (see Huh and Rusmevichientong (2009), Huh et al. (2009, 2011), Besbes and Muharremoglu (2013), Shi et al. (2016), Chen et al. (2015b)).

It should be noted that, even with complete information about the demand distribution *a priori*, the (clairvoyant) optimal policy for perishable inventory systems does not have any simple structure (see Nahmias (1975) and Fries (1975)), and computing the exact optimal policy is intractable using brute-force dynamic programming. Nandakumar and Morton (1993) studied these systems and noted that *"Since base stock policies are easier to implement and widely used in practice, the interest quickly turned to analyzing such policies for this problem"*. Indeed, a number of authors have investigated the performance of base-stock, or fixed critical number, policies for perishable inventory systems. For example, Cooper (2001) commented on this stream of research that *"The complexity of optimal policies, as well as the difficulties involved in computing them, has led many authors to analyze heuristic methods for controlling perishable inventories. One such method, proposed by Chazan and Gal (1977), Cohen (1976), Nahmias (1976), is the fixed-critical number order policy, in which orders are placed so that the total amount of inventory is the same at the end of each time period, regardless of the ages of the inventory. Computational studies by Nahmias (1976, 1977) and Nandakumar and Morton (1993) show that under a variety of different assumptions, such fixed-critical-number policies can be quite good when compared with other methods, as well as to optimal policies. In addition, these policies provide a good baseline against when other types of policies can be compared."* Further theoretical and computational evidences on the effectiveness of base-stock policies have been reported in Nahmias (1978), Deniz et al. (2010) and Cooper (2001).

These studies motivate us to *develop learning algorithms to find the best base-stock policies* for perishable inventory systems. Since the best base-stock policy performs very close to the true optimal policy, we shall use the long-run average cost of the clairvoyant best base-stock policy (had the distribution been known *a priori*) as our benchmark. This choice of benchmark is similar in spirit to Huh et al. (2009), which finds the best base-stock policy for the non-perishable lost-sales inventory control problem with positive lead times.

We aim to develop a nonparametric closed-loop control policy $\pi(S_t)$ for computing a *period-dependent* base-stock level S_t in each period t with unknown demand distribution *a priori* and censored demand information. Had the firm known the underlying demand distribution *a priori*, there exists a clairvoyant optimal base-stock policy $\pi(S^*)$. We measure the performance of our policy $\pi(S_t)$ through the notion of *regret*, the difference between the T-period average cost of running our policy $\pi(S_t)$ and the long-run average cost of the clairvoyant optimal base-stock policy $\pi(S^*)$. The main research question is to devise an effective nonparametric learning policy $\pi(S_t)$ that drives the average regret per period to zero with a provable (and tight) convergence rate.

1.2. Main Results and Contributions

Convexity result for base-stock policies. To facilitate the design of learning algorithms, we first prove a key structural result for perishable inventory systems operating under base-stock policies. More specifically, we show in Theorem 1 that the T-period total holding, lost-sales and outdating cost is convex in the base-stock level along every sample path for any $T \geq 1$. Our approach is linear programming (LP) based, which is similar in spirit to the one used in Janakiraman and Roundy (2004) for establishing a convexity result for non-perishable inventory systems with lost-sales and positive lead times.

Learning algorithm and regret analysis. The main contribution of this paper is to develop the first nonparametric learning algorithm, called the cycle-update policy (CUP for short), for finding the optimal base-stock policy in periodic-review perishable inventory systems with lost-sales and censored demand. We show in Theorem 2 that the average regret per period of CUP converges to zero at the rate of $O(1/\sqrt{T})$, which is theoretically the best possible. Our CUP algorithm belongs to the broad class of stochastic gradient descent algorithms. There are, however, several points of departure from the existing literature (see our literature review):

(a) First, our CUP algorithm is designed on carefully designed cycles, and it updates base-stock level in each cycle (rather than each period as done in other learning algorithms). The cycles are defined using the successive periods in which stockouts occur in the CUP system. These cycles are not *a priori* fixed but are sequentially triggered as demand realizes over time. Since the number of periods in each cycle is random, we have a random number of cycles, leading to technical difficulties that do not exist for the standard online optimization problems.

(b) Second, when we update base-stock level at the beginning of a cycle, computing a valid sample-path subgradient for the total costs accrued during the preceding cycle is a non-trivial task. The difficulty is caused by the dependence between the base-stock level and the amount of outdating units during a cycle. We develop a subroutine in §4.2 and show in Proposition 2 that it outputs a correct subgradient of total costs of a cycle with respect to the base-stock level.

(c) Since CUP updates base-stock levels in cycles, in the regret analysis, we need to compare and bound the total costs within a cycle between CUP and the clairvoyant optimal policy $\pi(S^*)$. However, one difficulty arises at the start of each cycle because the two policies considered have different inventory age distributions. In particular, CUP has zero initial inventory (due to a lost-sales event in the preceding period) and therefore all the inventory units ordered in that period are brand new with remaining lifetime m. But the age distribution of the optimal base-stock policy can be very different. To tackle this difficulty, we introduce a new bridging policy, called the replacement of old inventories (ROI for short), between CUP and the optimal base-stock policy. For each sample path, similar to the optimal base-stock policy, the bridging policy ROI uses S^* as its base-stock level, but at the beginning of each cycle, ROI replaces all its inventory units (regardless of their ages) with brand new inventory units (thus all having remaining lifetime m). We establish in Proposition 4 that for each sample path, the total cost incurred by ROI provides a lower bound on the total cost incurred by the optimal base-stock policy $\pi(S^*)$, which is a crucial intermediate step for the regret analysis.

1.3. Relevant Literature

Perishable inventory systems. With complete distributional information of demand, Nahmias (1975) and Fries (1975) studied the optimal policy for the general lifetime problem with i.i.d. demands, in a backlogging model and a lost-sales model, respectively. They showed that the optimal ordering policy depends on both the age distribution of the current inventory and the remaining period. We refer readers to Cooper (2001), Karaesmen et al. (2011) and Nahmias (2011) for a comprehensive overview on this topic. More recently, Chen et al. (2014) and Li and Yu (2014) derived new structural properties of optimal policies using the concepts of $L^\natural$-convexity and multimodularity, respectively. In parallel, Chao et al. (2015, 2018) and Zhang et al. (2016b) developed a series of approximation algorithms to compute provably near-optimal solutions for such systems. This paper differs from the above literature by not assuming the demand distribution *a priori*, and focuses on a joint learning and optimization problem under censored demand information.

Nonparametric algorithms for inventory models. A number of papers have been published on nonparametric algorithms for non-perishable inventory systems. Burnetas and Smith (2000) developed a learning algorithm for the repeated newsvendor problem with pricing. Huh and Rusmevichientong (2009) proposed a gradient descent based algorithm for lost-sales systems with censored demand. Besbes and Muharremoglu (2013) examined the discrete demand case and showed that active exploration is needed. Huh et al. (2011) applied the concept of Kaplan-Meier estimator to devise another data-driven algorithm for censored demand. Shi et al. (2016) proposed an algorithm for multi-product systems under a warehouse-capacity constraint. Chen et al. (2015a,b) proposed algorithms for the joint pricing and inventory control problem with backorders and lost-sales, respectively. Another popular nonparametric approach in the inventory literature is sample average approximation (SAA) (e.g., Kleywegt et al. (2002), Levi et al. (2007, 2015)) which uses the empirical distribution formed by *uncensored* samples drawn from the true distribution. Concave adaptive value estimation (e.g., Godfrey and Powell (2001), Powell et al. (2004)) successively approximates the objective cost function with a sequence of piecewise linear functions. The work closest to ours is perhaps Huh et al. (2009) that proposed a learning algorithm for finding the optimal base-stock policy in a lost-sales system with positive lead times. However, their system does not account for perishability, and the algorithms developed are based on different approaches.

1.4. General Notation

For any real numbers x and y, we denote $x^+ = \max\{x, 0\}$, $x \vee y = \max\{x, y\}$, and $x \wedge y = \min\{x, y\}$. The indicator function $\mathbb{1}(A)$ takes value 1 if A is true and 0 otherwise, and " $:=$ " stands for "defined as". We use LHS and RHS as abbreviations for "left-hand side" and "right-hand side", respectively. The projection function is defined as $\mathbf{P}_{[a,b]}(x) = \min[b, \max(x, a)]$ for any real numbers x, a, and b.

2. Perishable Inventory Systems with Censored Demand

We formally describe the stochastic periodic-review perishable inventory system with censored demand. The product lifetime m is known and fixed, i.e., items perish after staying in inventory for m periods if not consumed. Let $t \in \{1, 2, \ldots\}$ represent the time period, which is indexed forward. We denote the demand in period t by D_t, and assume that D_t, $t = 1, \ldots, T$, are i.i.d. continuous random variables across periods. Our model allows for both continuous and discrete demand distributions, though we focus on the continuous case (see Section 7 for a discussion on discrete demand). Contrary to the classical formulation, the firm has no prior knowledge about the true underlying demand distribution *a priori*, but can observe sales (i.e., censored demand), and make adaptive inventory decisions based on the available information.

System dynamics. Since any inventory unit that stays in the system for m periods without meeting the demand expires and exits the system, we use a state vector $\mathbf{x}_t$ to keep track of the inventory age information at the beginning of any period t (before ordering), i.e., $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,m-1})$, where $x_{t,i}$ is the on-hand inventory level of product whose remaining lifetime is *no more* than i periods, $i = 1, \ldots, m-1$. It is clear that $x_{t,m-1}$ is the total on-hand inventory level (*before* replenishment) in period t. We let q_t be the order quantity in period t, which can be any nonnegative real number. For notational convenience, we use $x_{t,m} = x_{t,m-1} + q_t$ to denote the total on-hand inventory level (*after* replenishment) in period t, which can be seen as our control variable. For any

FIFO issuance policy π, the sequence of events in each period t, $t = 1, 2, \ldots$, is as follows. (Note that there are many terms that depend on π and the sample path ω, such as $q_t^\pi, \mathbf{x}_t^\pi, o_t^\pi$, but for brevity, we shall make the dependency implicit.)

(a) At the beginning of each period t, the firm observes the starting inventory vector $\mathbf{x}_t$.

(b) The firm makes a replenishment decision $q_t \geq 0$ in period t, and the replenishment order arrives instantaneously. (Note that the zero lead time assumption is predominant in perishable inventory literature, see Nahmias (2011) and Karaesmen et al. (2011)). The total on-hand inventory level (after receiving the order q_t) is $x_{t,m} = x_{t,m-1} + q_t$.

(c) The random demand D_t is realized (denote its realization by d_t) and satisfied to the maximum extent under FIFO (i.e., the oldest inventory meets demand first). Under censored demand, the firm does not observe the realized demand d_t but observes the sales $\min(d_t, x_{t,m})$ only.

(d) At the end of the period, all the outstanding inventories incur a unit holding cost h and all the unsatisfied demands are *lost* with a unit lost-sales penalty cost p. Note that the lost-sales cost is unobservable in the event of stockout, due to demand censoring. Finally, all the inventories that have stayed in the system for m periods expire with unit outdating cost θ. Following the convention by Nahmias (1975), we assume the inventory units that perish at the end of this period also incur a holding cost. As a result, the period-t cost C_t^π is realized to be

$$h(x_{t,m} - d_t)^+ + p(d_t - x_{t,m})^+ + \theta(x_{t,1} - d_t)^+, \tag{1}$$

where the third term $o_t := (x_{t,1} - d_t)^+$ is the outdating inventory in period t. Note that the lost-sales $(d_t - x_{t,m})^+$ is unobservable due to demand censoring. We assume without loss of generality that the unit purchasing cost is zero (see a detailed cost transformation in Chao et al. (2015)).

(e) The system proceeds to period $t+1$ with $\mathbf{x}_{t+1}$ given by

$$x_{t+1,j} = (x_{t,j+1} - d_t - o_t)^+ = \Big(x_{t,j+1} - d_t - (x_{t,1} - d_t)^+\Big)^+, \quad \text{for } 1 \leq j \leq m-1. \tag{2}$$

The class of base-stock policies. Even with complete information about the demand distribution *a priori*, it is well-known that the (clairvoyant) optimal policy for perishable inventory systems is extremely complicated (see Karaesmen et al. (2011), Nahmias (2011)), and computing the exact optimal policy is intractable using brute-force dynamic programming. However, it has been shown in the literature that the class of *base-stock policies* has near-optimal computational performance (see Cooper (2001) and the detailed discussion in §1). Hence, in this paper, we focus our attention to find the best base-stock policy. Recall that under a base-stock policy of level S, the total inventory level at the beginning of each period is always raised to S, i.e., for any period t we have $q_t = (S - x_{t,m-1})^+$. We assume that the system is initially empty, i.e., $\mathbf{x}_1 = (0, \ldots, 0)$.

Without prior knowledge about the demand distribution, an admissible or feasible base-stock policy $\pi(S_t)$ is represented by a sequence of *period-dependent* order-up-to levels, $\{S_t, t \geq 1\}$, where

S_t depends only on the sales and decisions made prior to time t, i.e, S_t is adapted to the filtration generated by $\{S_s, \min\{S_s, D_s\} : s = 1, \ldots, t-1\}$ under censored demand. Focusing on this class of policies, we wish to develop a nonparametric adaptive inventory control policy $\pi(S_t)$ so that its average cost per period converges to that of the (clairvoyant) optimal base-stock policy, i.e.,

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi(S_t)}\right] = \inf_S \left\{\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi(S)}\right]\right\}, \tag{3}$$

where our adaptive policy $\pi(S_t)$ on the LHS of (3) is constructed under unknown demand distribution *a priori* and censored demand information while the (clairvoyant) optimal base-stock policy $\pi(S^*)$ on the RHS of (3) is constructed under known demand distribution *a priori*. We will also find the rate at which the average cost of policy $\pi(S_t)$ converges to that of $\pi(S^*)$.

3. Convexity for Base-Stock Policies

Assuming complete distributional information of demand, in this section we analyze the perishable inventory system operating under a constant base-stock policy $\pi(S)$. An important question is whether the total expected cost from period 1 to T is convex in the base-stock level S for any $T \geq 1$. The answer is affirmative and we shall provide an LP-based proof, which is similar in spirit to the one developed by Janakiraman and Roundy (2004) that proves the convexity of total cost in the base-stock level in a non-perishable inventory system with lost-sales and positive lead times.

Theorem 1. *For the perishable inventory systems operating under a base-stock policy $\pi(S)$, for any realization of demand $\omega = (d_1, d_2, \ldots)$, the T-period total cost is convex in S for any $T \geq 1$.*

The proof of Theorem 1 also suggests an interesting relationship between the optimal base-stock level S^* for the perishable inventory system and the optimal base-stock level $\tilde{S}^*$ for the counterpart of nonperishable inventory systems. Since the optimal base-stock level for non-perishable periodic-review inventory system has a closed form solution (i.e., the newsvendor quantile solution), this result gives an upper bound for the optimal base-stock level of perishable inventory system.

Proposition 1. *Consider a perishable inventory system and its counterpart of non-perishable inventory system with infinite lifetimes, under the same initial conditions, cost parameters and demand distributions. Denote the optimal base-stock levels for the perishable and nonperishable inventory systems by S^* and $\tilde{S}^*$, respectively. Then we have $S^* \leq \tilde{S}^*$.*

4. Nonparametric Algorithm: Cycle-Update Policy (CUP)

Not knowing the true underlying demand distribution D_t *a priori*, our objective is to find a provably good adaptive data-driven algorithm for inventory control such that its total expected system cost is close to that of the clairvoyant optimal base-stock policy. In the following, we present a *Cycle-Update Policy* (CUP for short) for the perishable inventory system with lost-sales and censored demand, which achieves the aforementioned objective.

We first make the following assumption about the (clairvoyant) optimal base-stock level S^*.

Assumption 1. *There is a known finite number $\bar{S}$ such that $S^* \leq \bar{S}$, and $\mathbb{P}(D_t \geq \bar{S}) > 0$.*

This is a mild and reasonable assumption since typically the firm has some idea about the maximum possible base-stock level. Similar assumptions have also appeared in Huh and Rusmevichientong (2009), Huh et al. (2009, 2011), Shi et al. (2016), Chen et al. (2015b) for other inventory systems. We remark that if the firm has some prior estimate of the demand distribution, the firm can readily compute $\tilde{S}^*$ for the counterpart (non-perishable) inventory system, which then serves an upper bound for S^* by Proposition 1.

We introduce the following notation to denote the total cost in periods $\{n_1, n_1+1, \ldots, n_2-1\}$ for any $1 \leq n_1 < n_2 \leq T+1$ operating under a base-stock policy S with *brand new* inventory level S in period n_1: $G(S,(n_1,n_2);\omega) = \sum_{t=n_1}^{n_2-1} C_t^{\pi(S)}(\omega)$, where $C_t^{\pi(S)}(\omega)$ is given in (1). Then, by Theorem 1, $G(S,(n_1,n_2);\omega)$ is convex in S for any n_1, n_2 on every sample path ω.

4.1. Cycle-Update Policy (CUP)

The key idea of our cycle-update policy (CUP) algorithm is to update the base-stock level every time the system experiences a stockout (i.e., zero inventory level is observed by the firm), and keeps the current base-stock level unchanged otherwise. Define the stockout period as the end of a cycle. Algorithm 1 below describes the algorithm, which calls to a detailed routine for computing a cycle subgradient described in §4.2, that is used to compute the base-stock level for a new cycle. We note that the idea of decomposing the planning horizon into updating cycles has been used in Huh et al. (2009) for a lost-sales non-perishable inventory model, but their approach and analyses are very different than ours.

In contrast to the existing literature on online convex optimization, the cycles in our algorithm are not *a priori* fixed or known, and they are triggered sequentially by lost-sales events as demands realize over time. Specifically, let τ_k be the beginning of k-th cycle for which CUP implements a newly computed base-stock level S_k, $k=1,\ldots$, the cycle ends the first time after τ_k that stockout occurs. That is, for each sample path $\omega = \{d_1, d_2, \ldots\}$, $\tau_1(\omega)=1$, and for $k \geq 1$,

$$\tau_{k+1}(\omega) = \inf\{t \geq \tau_k(\omega)+1 : x_{t,m-1}(\omega) = 0\}.$$

The k-th cycle cost of the CUP is $G(S_k,(\tau_k,\tau_{k+1});\omega)$. Note that $\tau_{k+1}-\tau_k$ is geometrically distributed with parameter $P(D \geq S_k)$, where D is a generic single-period demand. In addition, τ_{k+1} is *not* independent of the costs $C_{\tau_k}^{\pi(S_k)}, \ldots, C_{\tau_{k+1}-1}^{\pi(S_k)}$ incurred in cycle k.

In what follows, *we let $\nabla_1 G(S_k,(n_1,n_2);\omega)$ denote the partial subderivative of $G(S_k,(n_1,n_2);\omega)$ with respect to S_k.*

Algorithm 1 Cycle-Update Policy (CUP)

Initialization. Set $\tau_1 = 1$, and the initial base-stock level S_1 for period 1 is arbitrarily chosen from $(0, \bar{S})$. Set $x_{1,m-1} = S_1$, and set the cycle counter to $k = 1$.

Main Step. For each period $t \geq 2$, repeat the following procedure:

Case 1: If the starting inventory level $x_{t,m-1} > 0$ (i.e., lost-sales did not occur in period $t-1$), then keep the same base-stock level as in period $t-1$, i.e., order up to S_k in period t so that $x_{t,m} = S_k$. Go to the next period.

Case 2: If the starting inventory level $x_{t,m-1} = 0$ (i.e., lost-sales occurred in period $t-1$), then set $\tau_{k+1} = t$ at the beginning of a new cycle $k+1$, and update the base-stock level S_{k+1} by

$$S_{k+1} = \mathbf{P}_{[0,\bar{S}]}\big(S_k - \eta_k \nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega)\big), \tag{4}$$

where the step-size $\eta_k = \gamma/\sqrt{k}$ for some positive constant γ, and $\nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega)$ is a subgradient of the k-th cycle cost with respect to S_k *fixing* τ_k and τ_{k+1}, which can be efficiently computed using a *subroutine* presented in §4.2 via (5). Order up to S_{k+1} for period t so that $x_{t,m} = S_{k+1}$, and set $k := k+1$. Go to the next period.

Remark 1. One important observation of our CUP algorithm is that the starting inventory level in each period is always below the base-stock level. This is because CUP updates the base-stock level only when the system becomes empty, and keeps the same base-stock level otherwise. This implies that CUP can always attain the desired base-stock level in each period. It shall be noted that many papers in the demand learning literature need to deal with the "overshooting" or "undershooting" issue of not being able to achieve the desired base-stock levels in some periods due to either positive inventory carry-over or capacity constraints (e.g., Huh and Rusmevichientong (2009), Huh et al. (2009), Shi et al. (2016), Chen et al. (2015b)). We resolve this issue by the algorithmic design of CUP, which greatly simplifies the performance analysis.

We further remark that, it is not possible to design cycles based on successive stockout events defined by $\{D \geq \bar{S}\}$. This is because the total inventory level in our system is less than or equal to $\bar{S}$, and the event $\{D \geq \bar{S}\}$ may not be unobservable in our system due to censored demand. This is why we design cycles using successive stockouts of CUP, which is always feasible under censored demand. Note that under our design, CUP starts a new cycle with empty initial inventory, but the optimal policy may start a new cycle with positive inventory consisting of different ages, which gives an extra layer of complexity when carrying out the performance analysis.

4.2. Computing Cycle Subgradient

The above CUP algorithm requires computing a valid sample-path subgradient of the total cost within a cycle with respect to S_k for every sample path ω. The cycle subgradient is the sum of the following two parts. It is important to note that, τ_{k+1} depends on S_k, however, we only compute the partial subderivative of cycle cost for fixed τ_k and τ_{k+1}.

Cycle subgradient of the holding and lost-sales cost. The computation of a subgradient (w.r.t. S_k) of the cycle holding and lost-sales cost is straightforward, and it is simply $h \cdot (\tau_{k+1} - \tau_k - 1) - p$ for cycle $k = 1, 2, \ldots$. This is because, by the definition of the k-th cycle, namely periods τ_k to $\tau_{k+1} - 1$, the CUP algorithm ends with positive inventory during the first $(\tau_{k+1} - \tau_k - 1)$ periods, and experiences a stockout in the last period $\tau_{k+1} - 1$.

Cycle subgradient of the outdating cost. The computation of a subgradient of cycle outdating cost is more involved (but can be efficiently computed). Focus on the k-th cycle, namely $\{\tau_k, \ldots, \tau_{k+1} - 1\}$, in which the base-stock level is kept at S_k, and objective is to compute a subgradient of the cycle outdating cost with respect to S_k. Let $u_{t,i}$ denote the left subderivative of inventory with remaining lifetime i with respect to S_k for fixed τ_k and τ_{k+1}.

In each period t, besides the inventory vector $\mathbf{x_t}$, where the newly received order q_t in period t is considered as inventory with remaining lifetime m, we will keep track of another m-dimensional vector $\mathbf{u}_t = (u_{t,1}, \ldots, u_{t,m})$, which represents the subderivatives of inventory levels of different remaining lifetimes with respect to S_k. Since the sum of all inventory units is S_k in each period t during cycle k, the sum of all the entries of $\mathbf{u_t}$ must be 1. In fact, it can be argued that $u_{t,i} \in \{0, 1\}$ for all $i = 1, \ldots, m$ (see the proof of Proposition 2). For notational convenience, we use $\mathbf{e}_i^m$ to denote an m-dimensional vector whose i-th entry is 1 and all other entries are 0. Then $\mathbf{u_t} \in \{\mathbf{e}_1^m, \ldots, \mathbf{e}_m^m\}$ for each period t.

The following subroutine specifies how $\mathbf{u_t}$ is updated during the k-th cycle, and it is used to determine the cycle subgradient of outdating cost used in our CUP algorithm.

Subroutine Computing the k-th Cycle Subgradient for Algorithm 1

Initialization: In period $t = \tau_k$, initialize $\mathbf{u_t} := \mathbf{e}_m^m$, and a counter $n = 0$.

Main Step: For each subsequent period $t = \tau_k + 1, \ldots, \tau_{k+1}$, suppose $\mathbf{u_{t-1}} := \mathbf{e}_i^m$ for some $i \in \{1, \ldots, m\}$.

Case 1: If no outdating occurs in period $t-1$, then let $j = \min\{\ell : x_{t,\ell} > 0\}$ denote the remaining lifetime of the oldest inventory in period t (after ordering), and set $\mathbf{u_t} := \mathbf{e}_{\max(i-1,j)}^m$.

Case 2: If outdating occurs in period $t-1$, then

(i) if $i = 1$, then set $\mathbf{u_t} := \mathbf{e}_m^m$, and set $n := n + 1$;

(ii) otherwise set $\mathbf{u_t} := \mathbf{e}_{i-1}^m$.

Proposition 2. *Let n be the output of the subroutine for cycle k, then a valid subgradient for the k-th cycle outdating cost is θn.*

The main idea underlying this subroutine is the following: We perturb the base-stock level S_k by an infinitesimal amount to $S_k+\delta$, and compute the additional amount of outdating inventory due to such a change within the cycle (see Figure 1 as an example). Call the two systems, with base-stock levels S_k and $S_k+\delta$, the original system and the perturbed system, respectively. By base-stock policy, the total inventory of different remaining lifetimes in the two systems are always S_k and $S_k+\delta$, respectively, in each period of cycle k. Because demand is continuous, when δ is very small, with probability nearly 1 the inventory levels of different remaining lifetimes in the two systems are identical except at one entry, at which the perturbed system has δ more units of inventory than the original system. The intuition is as follows: Suppose these two systems differ in two entries in period t for the first time, then the demand in the previous period $t-1$ has to be strictly between $x_{t-1,1}$ and $x_{t-1,1}+\delta$, which has small probability of happening when δ is small. As a result, with high probability the two systems will differ only in one entry. By keeping track of the extra inventory level δ, which is precisely the unit vector $\mathbf{u_t}$, the subroutine allows us to compute how much more outdating inventory in the perturbed system than in the original system, and it is exactly $n\delta$ if n is the output of the subroutine.

Combining Proposition 2 and the cycle subgradient of the holding and lost-sales cost, we obtain the subgradient of the k-th cycle total cost as

$$\nabla_1 G(S_k,(\tau_k,\tau_{k+1});\omega)=\theta n+h\cdot(\tau_{k+1}-\tau_k-1)-p. \tag{5}$$

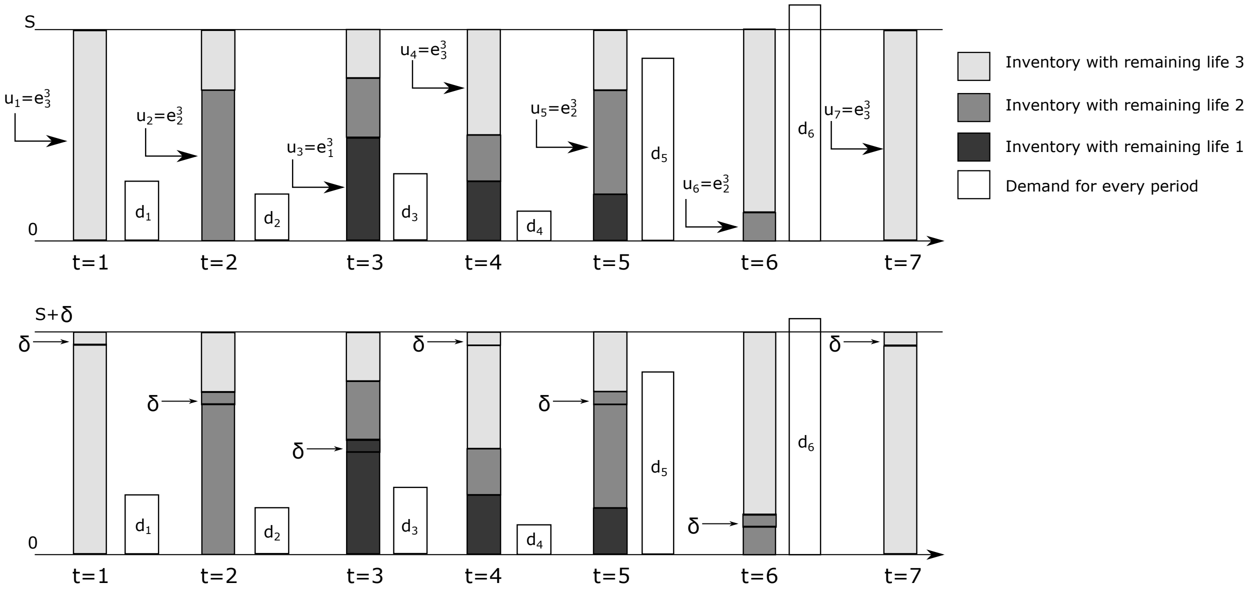

Figure 1 An example of how $\mathbf{u_t}$ is updated, with $m=3$.

Example 1. To better understand the above subroutine, we use a concrete example with $m = 3$ to illustrate how $\mathbf{u_t}$ is updated during a cycle. This example is designed to cover all possible scenarios. In the upper portion of Figure 1, we keep track of $\mathbf{u}_t$ for every period t, while in the lower portion of Figure 1, we perturb the base-stock level by a small amount δ. We can see that after this perturbation, there is always an additional δ amount of inventory in each period t, and more importantly, the exact position (or remaining lifetime) of this δ amount of inventory is tracked using $\mathbf{u}_t$ for every period t.

In this example, there is no outdating in periods $1, 2, 5, 6$. As a result, when we update $\mathbf{u_t}$ for $t = 2, 3, 6, 7$, we identify the remaining lifetime of the oldest on-hand inventory unit in period t, and they are $2, 1, 2, 3$, respectively. We then update $\mathbf{u_t} = \mathbf{e}^m_{\max(i-1,j)}$ as shown in Figure 1. Outdating happens in periods 3 and 4, and we need to check the events $\{\mathbf{u_3} = \mathbf{e}^3_1\}$ and $\{\mathbf{u_4} = \mathbf{e}^3_1\}$. It turns out that the event $\{\mathbf{u_3} = \mathbf{e}^3_1\}$ is true but the event $\{\mathbf{u_4} = \mathbf{e}^3_1\}$ is false. Hence, the δ-outdating event happens only once during this cycle, and thus the cycle subgradient for the outdating cost is θ. □

Remark 2. One may think that a subgradient of the total outdating cost within a cycle is simply θ times the number of outdating periods within a cycle. Example 1 above clearly shows that this is incorrect. As illustrated in Figure 1, when we raise the base-stock level by δ, the outdating amount in period 3 increases by δ, but the outdating amount in period 4 stays unchanged (which in fact equals $d_1 - d_4$ in both cases). In this example, the naive way of computing the cycle subgradient of total outdating cost gives us 2θ, but in fact it should be θ.

4.3. Numerical Experiments

The performance of CUP is measured by the percentage of total T-period cost increase compared with that of the (clairvoyant) optimal base-stock policy. We tested two demand distributions: uniform on $[0, 100]$ and truncated normal on $[0, 100]$ with mean 50 and standard deviation 25. The cost parameters are $h = 1, \theta = 5$ and $p \in \{5, 10\}$. For CUP, we set $\bar{S} = 95$, and considered two starting target inventory levels $S_1 = 0$ and $S_1 = 50$, and two step-sizes $\gamma = 1$ and $\gamma = 2$. Each instance is run 5000 times to compute the average costs of CUP and the (clairvoyant) optimal base-stock policy. The performance of CUP for these testing instances are provided in Table 1. For simplicity, we only display the results for $T = \{50, 200, 500, 1000, 2000\}$. From Table 1, we can see that for $S_1 = 0$, the step-size $\gamma = 2$ generally performs better, whereas for $S_1 = 50$, $\gamma = 1$ generally performs better. Moreover, CUP converges faster when the starting target inventory is 50 than when it is 0, this is intuitive as 50 is closer to the best base-stock level under all cases. This suggests that, although the result on convergence rate holds under all initial target base-stock levels, in practice the firm should make use of available information to determine a better starting point to improve the performance of the algorithm.

	S_1	γ	$T=50$	200	500	1000	2000
Uniform, $p=5$	0	1	159.7%	57.2%	23.6%	11.8%	5.9%
		2	70.7%	19.1%	8.1%	4.3%	2.3%
	50	1	16.3%	5.1%	2.2%	1.2%	0.6%
		2	8.8%	3.6%	2.0%	1.2%	0.7%
Uniform, $p=10$	0	1	158.62%	42.67%	17.61%	9.14%	4.80%
		2	63.02%	19.00%	9.23%	5.45%	3.30%
	50	1	22.72%	7.11%	3.55%	2.14%	1.31%
		2	13.81%	8.29%	5.09%	3.44%	2.29%
Normal, $p=5$	0	1	204.51%	62.31%	25.31%	12.75%	6.44%
		2	81.10%	21.53%	9.23%	4.94%	2.68%
	50	1	11.64%	3.71%	1.76%	1.01%	0.58%
		2	7.46%	3.68%	2.18%	1.44%	0.94%
Normal, $p=10$	0	1	164.84%	43.17%	17.94%	9.37%	4.95%
		2	67.87%	22.13%	11.39%	6.82%	4.11%
	50	1	16.32%	5.98%	3.29%	2.10%	1.34%
		2	15.29%	13.24%	8.48%	5.40%	3.41%

Table 1 Percentage of total expected cost increase of CUP under different problem instances

5. Performance Analysis of CUP

Given a sample path $\omega = \{d_1, d_2, \ldots,\}$ of demand process, the T-period regret of our nonparametric adaptive inventory policy CUP is defined as the difference between the clairvoyant optimal cost (given the demand distribution *a priori*) and the cost incurred by CUP (which learns the demand distribution over time under censored demand information) over T periods. More precisely,

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

where S_t is the base-stock level prescribed by our nonparametric (closed-loop) algorithm CUP, and S^* is the (clairvoyant) optimal base-stock level defined in (3). The average regret of CUP is $\mathbb{E}[\mathcal{R}_T^{\mathbf{CUP}}]$, and the average regret per period is defined as $\mathbb{E}[\mathcal{R}_T^{\mathbf{CUP}}]/T$.

Theorem 2. *Suppose Assumption 1 holds. Then, there exists some positive constant K_1, such that for each problem instance of the perishable inventory system described in §2, the expected cumulative regret of the cycle-update policy (CUP) satisfies*

$$\mathbb{E}\left[\mathcal{R}_T^{\mathbf{CUP}}\right] \le K_1\sqrt{T}, \qquad \textit{for all } T \ge 1.$$

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

The proof of Theorem 2 is based on online convex optimization and general stochastic approximation, see e.g., Zinkevich (2003), Nemirovski et al. (2009), Duchi and Singer (2009), and Hazan (2016). However, there are two key differences between our approach and those in the literature.

The first difference is that the CUP algorithm contains cycles whose lengths are *a priori* random and potentially unbounded. The second difference is that it is difficult to directly compare the cost difference between CUP and the optimal policy, since the two policies considered have different inventory age distributions at the start of each cycle. Thus, we construct a bridging policy termed ROI to resolve this issue (see §5.1 for the detailed construction). We are able to bound the difference between the total costs of CUP and ROI, and between the costs of ROI and the optimal policy, that are used to complete the proof of Theorem 2.

Remark 3. When $\mathbb{P}(D_t \geq \bar{S}) := \mu > 0$ is known *a priori*, the constant coefficient K_1 in the regret bound can be minimized by choosing $\gamma = \mu\bar{S}/\left(\max(h+\theta,p)\cdot\sqrt{4-2\mu}\right)$ in the algorithm, which gives $K_1 = \max(h+\theta,p)\cdot\bar{S}\cdot\sqrt{4-2\mu}/\mu$. We refer the reader to the Electronic Companion for more details.

It can be shown that the square-root regret rate is tight, which is formally stated as below.

Proposition 3. *Suppose $T > 2$. There exist problem instances such that the expected cumulative regret for any learning algorithm is lower bounded by $\Omega(\sqrt{T})$.*

The proof of this proposition is constructed based on the discrete demand example by Besbes and Muharremoglu (2013), which shows that even with uncensored demand, the lower bound of any learning algorithms for the repeated newsvendor problem (that is a special case of our model with $m = 1$) is $\Omega(\sqrt{T})$. A slightly modified continuous demand example can be found in Zhang et al. (2016a). We therefore omit the proof.

5.1. A Bridging Policy – Replacement of Old Inventories (ROI)

Our strategy to prove Theorem 2 is to compare and bound the difference between the k-th ($k = 1, 2, \ldots$) cycle total costs of CUP and the clairvoyant optimal policy. A difficulty arising at the start of each cycle τ_k is that, the CUP algorithm updates the base-stock level to S_k and order up to it, and by the construction of CUP, the system is empty at the beginning of τ_k. Therefore, all the S_k inventory units at the beginning of τ_k (after replenishment) are new with remaining lifetime m. However, in the system operating under the (clairvoyant) optimal inventory, the age distribution of the S^* inventory units at the beginning of τ_k is unknown and can be arbitrary. This creates difficulty in comparing the cost difference between these policies in cycle k (between periods τ_k and $\tau_{k+1} - 1$). A key step in the proof of Theorem 2 is to resolve this issue.

To circumvent this difficulty, we introduce a bridging policy, called the replacement of old inventories (ROI for short), between CUP and the optimal base-stock policy $\pi(S^*)$. For each sample path, similar to the optimal policy, the bridging policy ROI uses S^* as its base-stock level. However, at the beginning of τ_k ($k = 1, 2, \ldots$), ROI replaces all its inventory units (regardless of their ages) with brand new inventory units with remaining lifetime m.

The next result shows that, for each sample path, the total cost incurred by ROI gives a lower bound on the total cost incurred by the optimal base-stock policy $\pi(S^*)$, and it provides a bridge in comparing the total costs of CUP and the optimal policy.

PROPOSITION 4. *For each problem instance of the perishable inventory system described in §2, given any sample path $\omega = \{d_1, d_2, \ldots\}$ and any $T \geq 1$, the total cost incurred by the bridging policy ROI is less than or equal to the total cost incurred by the optimal base-stock policy $\pi(S^*)$.*

6. Strongly Convex Extension

We extend our algorithm to the strongly convex case with continuous demand, and obtain an improved regret rate. A differentiable function $g(\cdot)$ defined on a convex set of $\mathbb{R}$ is called strongly convex with parameter $\lambda > 0$ (Hazan (2016)), if $g(y) \geq g(x) + \nabla g(x)(y-x) + \frac{\lambda}{2}\big(y-x\big)^2$ for all x, y.

ASSUMPTION 2. *There exist three known finite numbers $\bar{S}$, $\underline{S}$ and λ, such that*

(i) $0 \leq \underline{S} < \bar{S}$, $\lambda > 0$,

(ii) $\underline{S} \leq S^* \leq \bar{S}$, *and* $\mathbb{P}(D_t \geq \bar{S}) > 0$, *and*

(iii) the probability density function $f(x)$ of single-period demand D satisfies $\inf_{x \in [\underline{S}, \bar{S}]} f(x) \geq \lambda$.

With the slightly stronger Assumption 2 (in place of Assumption 1), we can show that the expected cost of a cycle is strongly convex in the base-stock level S, and a modified version of CUP achieves a logarithmic regret rate, i.e., its average regret converges to zero at rate $O\big((\log T)/T\big)$.

THEOREM 3. *For perishable inventory system, we modify CUP or Algorithm 1 as follows:*

(1) Use the projection operator $\mathbf{P}_{[\underline{S}, \bar{S}]}$, *instead of using* $\mathbf{P}_{[0, \bar{S}]}$.

(2) Change the step-size to $\eta_k = \left(\frac{1}{\lambda(h+p)}\right)\frac{1}{k}$, $k = 1, 2, \ldots$.

Then under Assumption 2, there exists some positive constant K_2, such that for any $T \geq 1$, the expected cumulative regret of CUP for any problem instance satisfies $\mathbb{E}\left[\mathcal{R}_T^{\mathbf{CUP}}\right] \leq K_2 \log T$.

For the general online convex optimization problem, the improvement to $O(\log T)$ for the strongly convex case is typically straightforward (see, e.g., Zinkevich (2003), Nemirovski et al. (2009), Huh and Rusmevichientong (2009), Hazan (2016)). However, this extension is rather non-trivial in our model. The key reason is that we only have that the expected (not sample-path) cycle cost function is strongly convex, and the CUP algorithm involves random cycles that correlate with the random cycle costs, which leads to some technical difficulties in developing the regret bound. Indeed, when we work with expected regret, the number of random cycles depends on CUP, which then correlates with its random cycle cost, and the standard argument based on Wald's Theorem does not work. To circumvent this technical issue, we "stretch" the time horizon from period T to the T-th cycle of CUP, then show that the cumulative regret over T periods is upper bounded by the cumulative regret over T cycles plus a constant, and study the regret of the T-cycle problem.

7. Concluding Remarks

In §2, we assume that the order quantity in each period can be any nonnegative real number. In practice, the set of possible ordering quantities may be constrained to a discrete set. When the demand distribution and the ordering quantities are both discrete, our proposed learning algorithm CUP can be randomized in a similar way as that in §3.4 of Huh and Rusmevichientong (2009), and the result would still hold if the firm knows some additional information termed *the lost-sales indicator*. We refer the reader to §3.4 of Huh and Rusmevichientong (2009) for more discussion.

We offer some additional comments on the role of Assumption 1 on the upper bound of optimal base-stock level. This assumption prevents the order-up-to level to be too high so that there is always a chance to stock out, which in turn ensures that the time to the next stockout is bounded by a geometric random variable. If this assumption is not satisfied, i.e., the $\bar{S}$ that satisfies the condition is not known to the decision maker, we propose to modify the CUP algorithm as follows. Let the cycle end when either a stockout occurs or the number of consecutive periods without stockout reaches a predetermined threshold. This allows us to terminate the cycle prematurely. (Note that the original CUP algorithm is a special case with the threshold being infinity). One issue created by this modification is that the system may not have empty initial inventory at the beginning of a cycle (due to premature termination). To resolve this issue, we need to introduce another bridging system that starts with empty initial inventory at the beginning of each cycle. The subgradient of cycle cost in this bridging system can be computed using the sales data of our system, which is then employed in the algorithm to update the base-stock level for the subsequent cycle. The threshold (for early termination of a cycle) should be increasing in the cycle index, which can be selected carefully to achieve a minimum regret rate.

We close this paper by pointing out three more directions for future research. First, extending the current model to incorporate positive lead times is an important direction. However, the argument used in Theorem 1 for proving the convexity of total cost with respect to base-stock level fails to work when the ordering lead time is positive, and one needs to resolve this issue before applying online convex optimization algorithms. Second, it would be interesting to extend the current model to accommodate nonstationary demands, e.g., the seasonal demand model analyzed in Huh and Rusmevichientong (2014), or the bounded variation model studied in Besbes et al. (2015). Third, one may also consider the counterpart models with stochastic lifetimes (which may require non-crossing assumptions). Developing learning algorithms and regret bounds for these models may require new approaches different than the ones used in this paper.

Acknowledgments

The authors thank the area editor, the anonymous associate editor, and the three anonymous referees for their constructive comments and suggestions, which helped 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., Y. Gur, A. J. Zeevi. 2015. Non-stationary stochastic optimization. *Operations Research* **63**(5) 1227–1244.

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

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

Chao, X., X. Gong, C. Shi, C. Yang, H. Zhang, S. X. Zhou. 2018. Approximation algorithms for capacitated perishable inventory systems with positive lead times. *Management Science*. Articles in Advance https://doi.org/10.1287/mnsc.2017.2886.

Chao, X., X. Gong, C. Shi, H. Zhang. 2015. Approximation algorithms for perishable inventory systems. *Operations Research* **63**(3) 585–601.

Chazan, D., S. Gal. 1977. A Markovian model for a perishable product inventory. *Management Science* **23**(5) 512–521.

Chen, B., X. Chao, H.-S. Ahn. 2015a. Coordinating pricing and inventory replenishment with nonparametric demand learning. Working paper, University of Michigan, Ann Arbor, MI. Available at http://ssrn.com/abstract=2694633.

Chen, B., X. Chao, C. Shi. 2015b. Nonparametric algorithms for joint pricing and inventory control with lost-sales and censored demand. Working paper, University of Michigan, Ann Arbor, MI. Available at http://ssrn.com/abstract=2700491.

Chen, X., Z. Pang, L. Pan. 2014. Coordinating inventory control and pricing strategies for perishable products. *Operations Research* **62**(2) 284–300.

Cohen, M. 1976. Analysis of single critical number ordering policies for perishable inventories. *Operations Research* **24**(4) 726–741.

Cooper, W. L. 2001. Pathwise properties and performance bounds for a perishable inventory system. *Operations Research* **49**(3) 455–466.

Deniz, B., I. Karaesmen, A. Scheller-Wolf. 2010. Managing perishables with substitution: Inventory issuance and replenishment heuristics. *Manufacturing & Service Operations Management* **12**(2) 319–329.

Duchi, J., Y. Singer. 2009. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research* **10** 2899–2934.

Fries, B. 1975. Optimal ordering policy for a perishable commodity with fixed lifetime. *Operational Research* **23**(1) 46–61.

Godfrey, G. A., W. B. Powell. 2001. An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science* **47**(8) 1101–1112.

Hazan, E. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization* **2**(3-4) 157–325.

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. H., P. Rusmevichientong. 2014. Online sequential optimization with biased gradients: Theory and applications to censored demand. *INFORMS Journal on Computing* **26**(1) 150–159.

Huh, W. H., P. Rusmevichientong, R. Levi, J. Orlin. 2011. Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator. *Operations Research* **59**(4) 929–941.

Huh, W. T., G. Janakiraman, J. A. Muckstadt, P. Rusmevichientong. 2009. 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.

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.

Karaesmen, I. Z., A. Scheller-Wolf, B. Deniz. 2011. Managing perishable and aging inventories: Review and future research directions. *International Series in Operations Research & Management Science* **151** 393–436.

Kleywegt, A. J., A. Shapiro, T. Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM J. on Optimization* **12**(2) 479–502.

Levi, R., G. Perakis, J. Uichanco. 2015. The data-driven newsvendor problem: New bounds and insights. *Operations Research* **63**(6) 1294–1306.

Levi, R., R. O. Roundy, D. B. Shmoys. 2007. Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* **32**(4) 821–839.

Li, Q., P. Yu. 2014. Multimodularity and its applications in three stochastic dynamic inventory problems. *Manufacturing & Service Operations Management* **16**(3) 455–463.

Nahmias, S. 1975. Optimal ordering policies for perishable inventory-II. *Operational Research* **23**(4) 735–749.

Nahmias, S. 1976. Myopic approximations for the perishable inventory problem. *Management Science* **22**(9) 1002–1008.

Nahmias, S. 1977. On ordering perishable inventory when both demand and lifetime are random. *Management Science* **24**(1) 82–90.

Nahmias, S. 1978. The fixed charge perishable inventory problem. *Operations Research* **26**(3) 464–481.

Nahmias, S. 2011. *Perishable Inventory Systems*, vol. 160. International Series in Operations Research & Management Science. Springer, New York, USA.

Nandakumar, P., T. E. Morton. 1993. Near myopic heuristics for the fixed-life perishability problem. *Management Science* **39**(12) 1490–1498.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization* **19**(4) 1574–1609.

Powell, W., A. Ruszczyński, H. Topaloglu. 2004. Learning algorithms for separable approximations of discrete stochastic optimization problems. *Mathematics of Operations Research* **29**(4) 814–836.

Prastacos, G. P. 1984. Blood inventory management: an overview of theory and practice. *Management Science* **30** 777–800.

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

Zhang, H., X. Chao, C. Shi. 2016a. Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. Working paper, University of Michigan, Ann Arbor, MI. Available at http://ssrn.com/abstract=2922820.

Zhang, H., C. Shi, X. Chao. 2016b. Approximation algorithms for perishable inventory systems with setup costs. *Operations Research* **64**(2) 432–440.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. Tom Fawcett, Nina Mishra, eds., *Proceedings of the 20th International Conference on Machine Learning (ICML)*. AAAI Press, Cambridge, MA, USA, 928–936.

Brief Bio:

Huanan Zhang is an assistant professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at the Pennsylvania State University. His primary research interest lies in stochastic optimization and online learning algorithms with applications to inventory and supply chain management, and revenue management.

Xiuli Chao is a professor in the Department of Industrial and Operations Engineering at the University of Michigan. His current research interests include stochastic modeling and analysis, inventory control, game applications in supply chains, and data-driven optimization.

Cong Shi is an assistant professor in the Department of Industrial and Operations Engineering at the University of Michigan. His research lies in stochastic optimization and online learning algorithms with applications to inventory and supply chain management, and revenue management.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Electronic Companion to "Perishable Inventory Systems: Convexity Results for Base-Stock Policies and Learning Algorithms under Censored Demand"

by Huanan Zhang, Xiuli Chao, and Cong Shi

Proof of Theorem 1

We shall prove a stronger result that, the total holding cost, the total lost-sales cost, and the total outdating cost are all convex in S. The total holding and lost-sales costs, when running a base-stock policy $\pi(S)$, are $h\sum_{t=1}^{T}(S-d_t)^+$ and $p\sum_{t=1}^{T}(d_t-S)^+$, respectively, and they are clearly convex in S. In the following, we prove that the total outdating cost is also convex in S.

First, we observe that under the base-stock policy $\pi(S)$ and zero lead time, the amount of outdating inventory units in any period for the lost-sales model is identical to that for the backlogging model. (In the backlogging model, the replenishment shall first fulfill the backlogged demand, and then raise the inventory to the desired target level.) This holds for any realization of demand process. It suffices to analyze the backlogging counterpart model for the remainder of this proof.

The backlogging perishable inventory system also starts with zero inventory. Under the base-stock policy $\pi(S)$, the on-hand inventory level (after ordering) is $x_{t,m}=S$ for all $t=1,\ldots,T$. For any given $d_1,\ldots,d_{T-1}$, we construct a linear program $\mathbf{LP}(S)$ as follows:

$$
\begin{aligned}
\min_{\{\mathbf{q},\mathbf{x}\}} \quad & \sum_{t=2}^{T} q_t && \text{(EC.1)}\\
\text{subject to} \quad & x_{1,i}=0, \quad i=1,\ldots,m-1, && \text{(EC.2)}\\
& x_{t,m}=S, \quad t=1,\ldots,T, && \text{(EC.3)}\\
& x_{t+1,i-1}=x_{t,i}-q_{t+1}, \quad t=1,\ldots,T-1,\ i=2,\ldots,m, && \text{(EC.4)}\\
& q_{t+1}\geq d_t, \quad t=1,\ldots,T-1, && \text{(EC.5)}\\
& q_{t+1}\geq x_{t,1}, \quad t=1,\ldots,T-1. && \text{(EC.6)}
\end{aligned}
$$

The decision variables are $\mathbf{q}=(q_t)_{t=2,\ldots,T}$, and $\mathbf{x}=(x_{t,i})_{t=1,\ldots,T,i=1,\ldots,m}$. For notational convenience, we denote the feasible region (EC.2–EC.6) by $\mathbf{\Gamma}(S)$.

We claim that, the unique feasible solution within $\mathbf{\Gamma}(S)$ that satisfies the system of equations

$$
q_t=\max(d_{t-1},x_{t-1,1}), \quad t=2,\ldots,T \qquad \text{(EC.7)}
$$

is an optimal solution to $\mathbf{LP}(S)$. Since for any feasible solution $(\mathbf{q},\mathbf{x})$ to $\mathbf{LP}(S)$, $\mathbf{x}$ is completely and uniquely determined by $\mathbf{q}$ using (EC.3) and (EC.4), we will focus on $\mathbf{q}$ in the remainder of the proof (while leaving $\mathbf{x}$ implicit). Note that the variable $\mathbf{x}$ in $\mathbf{LP}(S)$ is slightly different from the state variable defined in the systems dynamics of the lost-sales model, as each entry $x_{t,i}$ could be negative in this LP. However, the solution $(\mathbf{q},\mathbf{x})$ that satisfies $\mathbf{\Gamma}(S)$ and (EC.7) (which will be shown to be unique) completely describes the evolution of a backlogging perishable inventory system operating under an order-up-to-S policy, and the positive part of $\mathbf{x}$ in this LP is the same as the inventory vector in the lost-sales system.

To prove the claim above, we first argue that for any given $d_1,\ldots,d_{T-1}$, there is a unique solution $\hat{\mathbf{q}}$ that satisfies $\mathbf{\Gamma}(S)$ and (EC.7). Combining (EC.3) and (EC.4) from $\mathbf{\Gamma}(S)$, we have

$$\hat{x}_{t,1} = S - \hat{q}_{t-m+2} - \hat{q}_{t-m+3} - \cdots - \hat{q}_t, \qquad t = m,\ldots,T. \tag{EC.8}$$

By (EC.7), $\hat{q}_2 = \max(d_1,\hat{x}_{1,1}) = \max(d_1,0) = d_1$, which is unique. For $t = 3,\ldots,T$, $\hat{q}_t = \max(d_{t-1},\hat{x}_{t-1,1})$ which is determined using only $\hat{q}_2,\ldots,\hat{q}_{t-1}$ due to (EC.8). Hence, the solution $\hat{\mathbf{q}} = (\hat{q}_2,\ldots,\hat{q}_T)$ can be sequentially and uniquely determined.

We proceed to prove that $\hat{\mathbf{q}}$ is also optimal by constructing this solution from an arbitrary optimal solution $\mathbf{q}^0$ in $T-2$ steps. (It is clear that $\mathbf{LP}(S)$ is bounded below by zero so optimal solutions must exist.) For notational convenience, we denote the solution after step k by $\mathbf{q}^k$ (while keeping its corresponding $\mathbf{x}^k$ implicit). In each step $k = 1,\ldots,T-2$, we keep $\mathbf{q}^k$ feasible without changing its objective value. We shall argue that $\mathbf{q}^{T-2} = \hat{\mathbf{q}}$, which has the desired property (EC.7).

In the first step, we carry out the following operations. If $q_2^0 = \max(d_1, x_{1,1}^0)$, then set $\mathbf{q}^1 = \mathbf{q}^0$. Otherwise, set

$$\begin{cases} q_2^1 = \max(d_1, x_{1,1}^0), \\ q_3^1 = q_3^0 + q_2^0 - \max(d_1, x_{1,1}^0), \\ q_j^1 = q_j^0 \ \text{, for } j = 4,\ldots,T. \end{cases}$$

Note that the corresponding $\mathbf{x}^1$ will also be determined by $\mathbf{q}^1$. It is clear that the objective value remains unchanged. Moreover, $\mathbf{q}^1$ is also feasible, since after this operation, $x_{1,1}^1 = 0$ is unchanged, $x_{2,1}^1$ is raised as much as q_3^1 is raised (hence keeping q_3^1 feasible), and $x_{t,1}^1$ is non-increasing for all $t = 3,\ldots,T-1$.

Then, in the subsequent steps $k = 2,\ldots,T-2$, we check if $q_{k+1}^{k-1} = \max(d_k, x_{k,1}^{k-1})$. If yes, then set $\mathbf{q}^k = \mathbf{q}^{k-1}$. Otherwise, set

$$\begin{cases} q_j^k = q_j^{k-1} \ \text{, for } j = 2,\ldots,k, \\ q_{k+1}^k = \max(d_k, x_{k,1}^{k-1}), \\ q_{k+2}^k = q_{k+1}^{k-1} + q_{k+1}^{k-1} - \max(d_k, x_{k,1}^{k-1}), \\ q_j^k = q_j^{k-1} \ \text{, for } j = k+3,\ldots,T. \end{cases}$$

By the identical argument, we can show that $\mathbf{q}^k$ is feasible and gives the same objective value as the previous solution $\mathbf{q}^{k-1}$.

After these $T-2$ steps, we have obtained a feasible $\mathbf{q}^{T-2}$ that satisfies (EC.7) for $t = 2,\ldots,T-1$. It remains to verify that $q_T^{T-2} = \max(d_{T-1}, x_{T-1,1}^{T-2})$. This holds because if otherwise $q_T^{T-2} >$

$\max(d_{T-1}, x^{T-2}_{T-1,1})$, then we could decrease q^{T-2}_T until the inequality becomes binding, which gives rise to a new feasible solution that has a strictly lower objective value than $\mathbf{q}^0$, contradicting to the assumption that $\mathbf{q}^0$ is an optimal solution. Hence, we have that $\mathbf{q}^{T-2}$ satisfies (EC.7) and is also optimal by the above construction argument. Furthermore, $\mathbf{q}^{T-2} = \hat{\mathbf{q}}$, since $\hat{\mathbf{q}}$ is the unique feasible solution that satisfies (EC.7). We have proven the claim.

Since the feasible region $\mathbf{\Gamma}(S)$ is a convex subset of the space of decision variable $\{\mathbf{q}, \mathbf{x}\}$ and the parameter S, it follows that the optimal objective value of the linear program $\mathbf{LP}(S)$ is convex in S. Because there is a unique optimal solution $\hat{\mathbf{q}}$ that satisfies $\mathbf{\Gamma}(S)$ and (EC.7) (completely describing the evolution of a backlogging perishable inventory system operating under an order-up-to-S policy), this shows that the total number of inventory units ordered $\sum_{t=1}^{T} \hat{q}_t$ under this order-up-to-S policy is convex in S. In addition, it is easy to see that when running an order-up-to-S policy, we have

$$\sum_{t=2}^{T} \hat{q}_t = \sum_{t=1}^{T-1} d_t + \sum_{t=1}^{T-1} o_t.$$

This shows that, for any sequence of demand realizations $d_1, \ldots, d_{T-1}$, the total outdating cost $\theta \sum_{t=1}^{T-1} o_t$ is convex in S. This completes the proof. **Q.E.D.**

Proof of Proposition 1

Denote the expected one-period holding and lost-sales cost by $L(S) = \mathbb{E}[h(S-d)^+ - p(d-S)^+]$. It is well-known that $\tilde{S}^*$ is the unique minimizer of $L(S)$, and $L'(\tilde{S}^*) = 0$. For the corresponding perishable inventory system, besides $L(S)$, the inventory system also incurs an outdating cost. Denote the expected long-run outdating cost by using a base-stock policy $\pi(S)$ by $A(S)$. By Theorem 1, $A(S)$ is convex in S. It is clear that $A(S)$ is increasing in S, and therefore $A'(S) \geq 0$ for any $S \geq 0$. Since $L'(S^*) + A'(S^*) = 0$, we must have $S^* \leq \tilde{S}^*$. This completes the proof. **Q.E.D.**

Proof of Proposition 2

The main idea underlying this subroutine is that when we perturb the current base-stock level S_k by an infinitesimal amount δ, we compute how many δ's outdate within the cycle (see Figure 1 as an example). To that end, we keep track of the additional δ units of inventory by using the auxiliary vector $\mathbf{u}_t$ in each period t. More precisely, if $\mathbf{u}_t = e^m_i$ for some $1 \leq i \leq m$, then there is an additional δ (as a result of raising the base-stock level to $S_k + \delta$) that have remaining lifetime i in the on-hand inventory in period t. Indeed, since the inventory levels of different remaining lifetimes add to $S_k + \delta$, the extra δ units have to appear somewhere. We can argue that the inventory levels of different remaining lifetimes in the two systems are identical except at one entry, at which the perturbed system has δ more units of inventory than the original system. This is because if these two systems differ in two entries is in period t for the first time, then the demand in the previous period $t-1$ has to be strictly between $x_{t-1,1}$ and $x_{t-1,1}+\delta$, which has near zero probability of happening with small enough δ. As a result, the two systems will only differ in one entry, and we need to prove that $\mathbf{u}_t$ is indeed keeping track of this difference.

We first consider the case where $x_{t,1} \neq d_t$ for the original system for all $t = 1, \ldots, T$.

We prove, by induction, that for any $t = \tau_k, \ldots, \tau_{k+1}-1$, if $\mathbf{u}_t = \mathbf{e}_i^m$ then after we raise the base-stock level S_k by an infinitesimal amount δ, the inventory with remaining lifetime i will increase by δ in period t, while the inventory levels with any other remaining lifetime remain unchanged.

First, following our subroutine, $u_{\tau_k} = e_m^m$. Recall that the system is empty at the beginning of period τ_k, hence the ordering quantity $q_{\tau_k} = S_k + \delta$ and inventory levels with remaining lifetime not equal to m are all 0. So the claim is true for $t = \tau_k$. Suppose that the claim has been shown for period $t-1$, and we want to prove that it also holds for $\mathbf{u}_t$.

For clarity, we consider two systems, one with base-stock level S_k, called the original system, and the other with base-stock level $S_k+\delta$, called by the perturbed system. By induction assumption suppose $\mathbf{u}_{t-1} = \mathbf{e}_i^m$ for some i, i.e., in period $t-1$ the perturbed system has the same inventory vector as the original system except for an additional δ with remaining lifetime i. We consider two cases separately below.

Case 1: If no outdating occurs in period $t-1$, then we have the following subcases.

a) If $i=1$ or $\mathbf{u}_{t-1} = \mathbf{e}_1^m$, then all units with remaining lifetime 1 are consumed by demand in both the original and the perturbed systems, and by FIFO, there will be δ extra units in the perturbed system with the oldest inventory in period t, i.e., $\mathbf{u}_t = \mathbf{e}_j^m$ where $j = \min\{\ell : x_{t,\ell} > 0\}$.

b) If $i > 1$, then the δ extra units in the perturbed system are either consumed or still in system in period t. In the former case, $\mathbf{u}_t = \mathbf{e}_j^m$ where $j = \min\{\ell : x_{t,\ell} > 0\}$ is the oldest inventory in period t; and in the latter case, $\mathbf{u}_t = \mathbf{e}_{i-1}^m$ since the δ units have one less period of remaining lifetime in period t.

For both subcases a) and b), we update the vector $\mathbf{u}_t = \mathbf{e}_{\max(i-1,j)}^m$, but no change is made on outdating quantities. This is consistent with Case 1 in the Subroutine.

Case 2: If outdating occurs in period $t-1$, then we have the following subcases.

c) If $i=1$ or $\mathbf{u}_{t-1} = \mathbf{e}_1^m$, then all units with remaining lifetime being 1 either satisfy demand or outdate in both systems, and because there are δ more units in the perturbed system with remaining lifetime being 1, δ more units of inventory outdate in the perturbed system, incurring extra outdating cost. In this subcase the number of outdated inventory is increased by δ in the perturbed system. At the beginning of period t, an ordering quantity of S_k and $S_k+\delta$ will be ordered, respectively, in the original and perturbed systems all having remaining lifetime m, hence $\mathbf{u}_t = \mathbf{e}_m^m$.

d) If $i > 1$, then the fact that there is oudatinng in period $t-1$ implies that the extra δ units with remaining lifetime i will still be in system and its remaining lifetime will be reduced to $i-1$ in period t. Thus, $\mathbf{u}_t = \mathbf{e}_{i-1}^m$.

This shows that in period t, $\mathbf{u}_t$ also represents the position of the extra δ inventory units in the perturbed system which completes the induction proof. This also means that, apart from finite non-differentiable points (where $x_{t,1} = d_t$ for some t), the subderivative of $x_{t,i}$ with respect to S_k is always 0 or 1 for all $t = 1, \dots, T$ and $i = 1, \dots, m$. Hence, every $x_{i,j}$ is just a piecewise linear function with respect to S_k, as it is clear that $x_{i,j}$ must be continuous with respect to S_k. We can then verify that at those non-differentiable points, $u_{t,i}$ represents the left subderivative of $x_{t,i}$, which is a valid subgradient.

The argument above shows that the output θn represents the subgradient of outdating cost with respect to S_k. This completes the proof of Proposition 2. **Q.E.D.**

Proof of Proposition 4

It suffices to show that for a given sample path ω and a given base-stock level S, an empty system with zero initial inventory gives the lowest total cost from period 1 to any period T, among all possible configurations of initial inventory that is less than or equal to S and with any age distributions.

We first make a simple yet important observation. That is, under a base-stock policy $\pi(S)$, the holding and lost-sales costs are independent of the initial inventory as well as its age distribution, and is only affected by demands and the given base-stock level S. Hence, the initial inventory and its age distribution only affect the outdating cost. To analyze the outdating cost, we consider a variant of the linear program (LP) introduced in the proof of Theorem 1. Similar to the proof of Theorem 1, we also consider a backlogging model, where the replenishment shall first fulfill the backlogged demand, and then raise the inventory to the desired target level. Note that the amount of outdating inventory units in any period for the lost-sales model is identical to that for the backlogging model.

Denote the initial inventory configuration by $\mathbf{a} = (a_1, \dots, a_{m-1})$, where $0 \le a_1 \le a_2 \le \cdots \le a_{m-1} \le S$. Note that a_i represents the initial inventory with remaining lifetime no more than i periods. For any given $d_1, \dots, d_{T-1}$ and any initial inventory configuration $\mathbf{a}$, we construct a linear program $\mathbf{LP}'(S, \mathbf{a})$ as follows.

$$\min_{\{\mathbf{q},\mathbf{x}\}} \quad \sum_{t=2}^{T} q_t \tag{EC.9}$$

$$\text{subject to} \quad x_{1,i} = a_i, \quad i = 1, \dots, m-1, \tag{EC.10}$$

$$x_{t,m} = S, \quad t = 1, \dots, T, \tag{EC.11}$$

$$x_{t+1,i-1} = x_{t,i} - q_{t+1}, \quad t = 1, \dots, T-1, \ i = 2, \dots, m, \tag{EC.12}$$

$$q_{t+1} \ge d_t, \quad t = 1 \dots, , T-1, \tag{EC.13}$$

$$q_{t+1} \ge x_{t,1}, \quad t = 1, \dots, T-1. \tag{EC.14}$$

The decision variables are $\mathbf{q} = (q_t)_{t=2,\dots,T}$, and $\mathbf{x} = (x_{t,i})_{t=1,\dots,T, i=1,\dots,m}$. For notational convenience, we denote the feasible region (EC.10–EC.14) by $\mathbf{\Gamma}'(S,\mathbf{a})$.

Let the unique solution satisfying (EC.7) and $\mathbf{\Gamma}'(S,\mathbf{0})$ be $\mathbf{q}^0$, and the optimal solution satisfying (EC.7) and $\mathbf{\Gamma}'(S,\mathbf{a})$ be $\hat{\mathbf{q}}$. Following an identical argument as that in the proof of Theorem 1, we have that $\mathbf{q}^0$ is optimal for $\mathbf{LP}'(S,\mathbf{0})$ and $\hat{\mathbf{q}}$ is optimal for $\mathbf{LP}'(S,\mathbf{a})$. Hence, to prove Proposition 4, it suffices to prove the following claim: The optimal objective value of $\mathbf{LP}'(S,\mathbf{0})$ is less than or equal to that of $\mathbf{LP}'(S,\mathbf{a})$ for any $\mathbf{a}$ with $0 \le a_1 \le a_2 \le \cdots \le a_{m-1} \le S$, i.e., the objective value of solution $\mathbf{q}^0$ is less than or equal to that of solution $\hat{\mathbf{q}}$.

We shall prove this claim by constructing $\hat{\mathbf{q}}$ from $\mathbf{q}^0$ in at most $T-1$ steps, where the objective value is kept non-decreasing in each step.

In the optimal solution $\mathbf{q}^0$ for $\mathbf{LP}'(S,\mathbf{0})$, we have $q^0_{t+1} = d_t$ for $t = 1,\dots,m-1$, since $x^0_{t,1} \le x^0_{1,t} = 0 \le d_t$ for $t = 1,\dots,m-1$. In fact, this means that with zero starting inventory, the system will not have any outdating units in the first $m-1$ periods. This solution $\mathbf{q}^0$ may not be feasible for $\mathbf{LP}'(S,\mathbf{a})$; however, if $\mathbf{q}^0$ turns out to be feasible, then it must be also optimal for $\mathbf{LP}'(S,\mathbf{a})$ because $\max(d_t, x_{t,1})$ remains unchanged for each $t = 1,\dots,,T-1$.

Now consider the more involved case where $\mathbf{q}^0$ is not feasible for $\mathbf{LP}'(S,\mathbf{a})$. In this case, we have $q^0_t \le \max(d_{t-1}, x^0_{t-1,1})$ for any $2 \le t \le T$. We shall construct $\hat{\mathbf{q}}$ from $\mathbf{q}^0$ in at most $T-1$ steps. Denote the solution after step k by $\mathbf{q}^k$ (while keeping its corresponding $\mathbf{x}^k$ implicit).

In the first step, if $q^0_2 = \max(d_1, x^0_{1,1})$, then we let $\mathbf{q}^1 = \mathbf{q}^0$ and proceed to the next step. Otherwise, we set

$$\begin{cases} q^1_2 = \max(d_1, x^0_{1,1}), \\ q^1_3 = q^0_3 + q^0_2 - \max(d_1, x^0_{1,1}), \\ q^1_j = q^0_j \text{ , for } j = 4,\dots,T. \end{cases}$$

It is clear that the objective value remains unchanged. Then in the subsequent step $k = 2,\dots,T-2$, if if $q^{k-1}_{k+1} = \max(d_k, x^{k-1}_{k,1})$, then we let $\mathbf{q}^k = \mathbf{q}^{k-1}$ and proceed to the next step. Otherwise, we set

$$\begin{cases} q^k_j = q^{k-1}_j \text{ , for } j = 2,\dots,k, \\ q^k_{k+1} = \max(d_k, x^{k-1}_{k,1}), \\ q^k_{k+2} = q^{k-1}_{k+2} + q^{k-1}_{k+1} - \max(d_k, x^{k-1}_{k,1}), \\ q^k_j = q^{k-1}_j \text{ , for } j = k+3,\dots,T. \end{cases}$$

Note that the objective value remains unchanged after these operations.

In the final step $k = T-1$, we set

$$\begin{cases} q^{T-1}_j = q^{T-2}_j \text{ , for } j = 2,\dots,T-1 \\ q^{T-1}_T = \max(d_{T-1}, x^{T-2}_{T-1,1}). \end{cases}$$

Then, we obtain $\mathbf{q}^{T-1} = \hat{\mathbf{q}}$, with unchanged objective value in the first $T-2$ steps and a possible increase in objective value in the final step $k = T-1$. This proves the claim and the desired result then follows. **Q.E.D.**

Proof of Theorem 2

Consider an arbitrary sample path $\omega = \{d_1, d_2, \ldots\}$ and a fixed T. We use $N = N(\omega)$ to denote the total number of cycles before period T, including possibly the last incomplete cycle. If the last cycle is not completed at T, then we truncate the cycle and also let $\tau_{N+1} - 1 = T$.

By Proposition 4, we know that the bridging policy ROI provides a lower bound on the (clairvoyant) optimal base-stock policy $\pi(S^*)$. We shall compare the costs between CUP and ROI.

$$\begin{aligned}
\mathcal{R}_T^{\mathbf{CUP}}(\omega) &= \sum_{t=1}^{T} \left(C_t^{\mathbf{CUP}}(\omega) - C_t^{\pi(S^*)}(\omega) \right) \\
&\le \sum_{t=1}^{T} \left(C_t^{\mathbf{CUP}}(\omega) - C_t^{\mathbf{ROI}}(\omega) \right) \\
&= \sum_{k=1}^{N} \sum_{i=\tau_k}^{\tau_{k+1}-1} \left(C_i^{\mathbf{CUP}}(\omega) - C_i^{\mathbf{ROI}}(\omega) \right) \\
&= \sum_{k=1}^{N} \big(G(S_k, (\tau_k, \tau_{k+1}); \omega) - G(S^*, (\tau_k, \tau_{k+1}); \omega) \big). \qquad \text{(EC.15)}
\end{aligned}$$

Note that CUP starts with brand new inventory units in period τ_k for all $k = 1, \ldots, N$, because CUP has experienced lost-sales in the previous period $\tau_k - 1$ (by the construction of CUP). Similarly, ROI starts with brand new inventory units in period τ_k as well for all $k = 1, \ldots, N$, as we have replaced all the old inventory units in these periods with new ones. By Theorem 1, we have that the cycle cost function $G(S_k, (n_1, n_2); \omega)$ is convex in S_k, thus

$$G(S_k, (\tau_k, \tau_{k+1}); \omega) - G(S^*, (\tau_k, \tau_{k+1}); \omega) \le \nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega) \, (S_k - S^*) \, . \qquad \text{(EC.16)}$$

Substituting (EC.16) into (EC.15) yields

$$\mathcal{R}_T^{\mathbf{CUP}}(\omega) \le \sum_{k=1}^{N} \nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega) \, (S_k - S^*) \, . \qquad \text{(EC.17)}$$

On the other hand, by our CUP algorithm (4), we have that for almost every ω and $k = 1, \ldots, N$,

$$(S_{k+1} - S^*)^2 \le (S_k - S^*)^2 - \frac{2\gamma}{\sqrt{k}} (S_k - S^*) \nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega) + \frac{\gamma^2 \left(\nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega) \right)^2}{k}. \qquad \text{(EC.18)}$$

Combining (EC.17) and (EC.18), and taking expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_T^{\mathbf{CUP}}\right] \leq \mathbb{E}\left[\sum_{k=1}^{N} \frac{\sqrt{k}}{2\gamma}\left((S_k - S^*)^2 - (S_{k+1} - S^*)^2\right)\right] \\ +\mathbb{E}\left[\sum_{k=1}^{N} \frac{\gamma}{2\sqrt{k}}\left(\nabla_1 G(S_k, (\tau_k, \tau_{k+1}))\right)^2\right]. \qquad \text{(EC.19)}\end{aligned}$$

We first analyze the first term on the RHS of (EC.19). By some simple algebra, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{N} \frac{\sqrt{k}}{2\gamma}\left((S_k - S^*)^2 - (S_{k+1} - S^*)^2\right)\right] \\ &\leq \frac{1}{\gamma}\mathbb{E}\left[\frac{1}{2}(S_1 - S^*)^2 - \frac{\sqrt{N}}{2}((S_{N+1} - S^*)^2\right] + \frac{1}{2\gamma}\mathbb{E}\left[\sum_{k=2}^{N}\left(\sqrt{k} - \sqrt{k-1}\right)(S_k - S^*)^2\right] \\ &\leq \frac{1}{\gamma}\bar{S}^2\left[\frac{1}{2} + \frac{1}{2}\sum_{k=2}^{N}\left(\sqrt{k} - \sqrt{k-1}\right)\right] = \frac{\sqrt{N}}{2\gamma}\bar{S}^2 \leq \frac{\sqrt{T}}{2\gamma}\bar{S}^2. \qquad \text{(EC.20)}\end{aligned}$$

Then we analyze the second term on the RHS of (EC.19). By (5), it is seen that for almost every ω, the absolute value of cycle subgradient $\nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega)$ is bounded above by $\max(h + \theta, p) \cdot (\tau_{k+1} - \tau_k)$. Noting that $\tau_{k+1} - \tau_k$ is a geometric random variable with parameter $\mathbb{P}(D \geq S_k)$. Letting $\mu \triangleq \mathbb{P}(D \geq \bar{S}) > 0$ (by Assumption 1), then we have $\mathbb{P}(D \geq S_k) \geq \mathbb{P}(D \geq \bar{S}) = \mu$. Denote $U \sim \text{Geo}(\mu)$ as a geometric random variable with parameter μ, then we can write

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{N} \frac{\gamma}{2\sqrt{k}}\left(\nabla_1 G(S_k, (\tau_k, \tau_{k+1}); \omega)\right)^2\right] &\leq \gamma\left(\max(h+\theta, p)\right)^2 \cdot \mathbb{E}[U^2] \cdot \sum_{k=1}^{T} \frac{1}{2\sqrt{k}} \\ &\leq \frac{\gamma(2-\mu)\left(\max(h+\theta, p)\right)^2}{\mu^2} \cdot \sqrt{T}, \qquad \text{(EC.21)}\end{aligned}$$

where the second inequality follows from $\sum_{k=1}^{T} 1/\sqrt{k} \leq 2\sqrt{T}$ and

$$\mathbb{E}[U^2] = \text{VAR}(U) + (\mathbb{E}[U])^2 = \frac{2-\mu}{\mu^2}.$$

Combining (EC.20) and (EC.21), we obtain

$$\mathbb{E}\left[\mathcal{R}_T^{\mathbf{CUP}}\right] \leq \frac{\sqrt{T}}{2\gamma}\bar{S}^2 + \frac{\gamma(2-\mu)\left(\max(h+\theta, p)\right)^2}{\mu^2} \cdot \sqrt{T} \leq K_1\sqrt{T} \qquad \text{(EC.22)}$$

for some positive constant K_1. This completes the proof of Theorem 2.

If the value of μ is known *a priori*, then the middle term in (EC.22) can be considered as a function of γ, and it is minimized when γ takes value

$$\gamma = \frac{\mu\bar{S}}{\max(h+\theta, p) \cdot \sqrt{4 - 2\mu}}.$$

This balances the two terms in the middle of (EC.22), and it gives a minimized value of K_1 as

$$K_1 = \frac{\max(h+\theta, p)\cdot \bar{S}\cdot\sqrt{4-2\mu}}{\mu}.$$

Q.E.D.

Proof of Theorem 3.

For convenience, let $\bar{\mu} := \mathbb{P}(D \geq \underline{S}) > 0$ and $\underline{\mu} := \mathbb{P}(D \geq \bar{S}) > 0$. Then $1 \geq \bar{\mu} \geq \underline{\mu} > 0$.

To facilitate our analysis, we call the event $\{D \geq \bar{S}\}$ as A, and define N_A^i as the period in which the event A occurs the i-th time. More precisely, given a sample path $\omega = \{d_1, d_2, \ldots\}$,

$$N_A^{i+1}(\omega) = \inf\left\{t \geq N_A^i(\omega) + 1 :\, d_t > \bar{S}\right\}, \qquad N_A^0(\omega) = 0.$$

Recall that given a sample path ω, $C_t^{\pi(S)}(\omega)$ is the cost incurred in period t when applying a base-stock level S in every period. We first present the following auxiliary result.

Lemma EC.1. *The (clairvoyant) optimal base-stock level satisfies*

$$S^* = \arg\min_S \mathbb{E}\left[\sum_{t=N_A^i+1}^{N_A^{i+1}} C_t^{\pi(S)}\right], \qquad i = 0, 1, 2, \ldots.$$

Proof. Since the inventory system becomes empty every time event A occurs, the costs between $N_A^i + 1$ and N_A^{i+1} are i.i.d. random variables, $i = 0, 1, \ldots$, where we define N_A^0 as 0. Hence, it suffices to prove

$$S^* = \arg\min_S \mathbb{E}\left[\sum_{t=1}^{N_A^1} C_t^{\pi(S)}\right].$$

Consider an arbitrary base-stock level S and an arbitrary sample path. For any $t \geq 1$, let $J(t) = \max\{k : N_A^k \leq t\}$ be the number of cycles completed by time t, then we have $N_A^{J(T)} \leq T \leq N_A^{J(T)+1}$. As the total cost is non-decreasing in the number of periods, we must have

$$\frac{\sum_{j=1}^{J(T)}\left[\sum_{t=N_A^{j-1}+1}^{N_A^j} C_t^{\pi(S)}\right]}{J(T)} \cdot \frac{J(T)}{T} \leq \frac{\sum_{t=1}^T C_t^{\pi(S)}}{T} \leq \frac{\sum_{j=1}^{J(T)+1}\left[\sum_{t=N_A^{j-1}+1}^{N_A^j} C_t^{\pi(S)}\right]}{J(T)+1} \cdot \frac{J(T)+1}{T}.$$

Since the cycles have i.i.d. length with geometric distribution of mean $1/\underline{\mu}$, it follows from renewal theory that $J(T)/T$, as well as $(J(T)+1)/T$, converge almost surely to $\underline{\mu}$. Moreover, since $\sum_{t=N_A^{j-1}+1}^{N_A^{j+1}} C_t^{\pi(S)}$ for $i = 1, 2, \ldots$ are also i.i.d., it follows from the Strong Law of Large Numbers that, with probability 1,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T C_t^{\pi(S)} = \underline{\mu} \cdot \mathbb{E}\left[\sum_{t=1}^{N_A^1} C_t^{\pi(S)}\right]. \tag{EC.23}$$

It can be seen that the LHS of (EC.23) is almost surely bounded by $(h+\theta)\cdot\bar{S}+p\cdot\frac{1}{T}\sum_{t=1}^{T}D_t$, which is integrable, thus applying Lebesgue's Dominated Convergence Theorem we obtain

$$\lim_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}C_t^{\pi(S)}\right]=\mathbb{E}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}C_t^{\pi(S)}\right]=\underline{\mu}\cdot\mathbb{E}\left[\sum_{t=1}^{N_A^1}C_t^{\pi(S)}\right]. \tag{EC.24}$$

Because S^* minimizes the first term of (EC.24), it also minimizes the third term. This completes the proof of Lemma EC.1 **Q.E.D.**

With Lemma EC.1, we are ready to prove Theorem 3.

Proof of Theorem 3. We define $b(t)$ as the first period after t that an event A occurs, i.e., $b(t)=\inf\left\{s\geq t:\, d_t>\bar{S}\right\}$. It is important to note that these stopping times $b(t)$'s are policy-independent. We further introduce the notation $l(T)$ to denote the end of T-th cycle of CUP, and it is clear that $l(T)\geq T$ almost surely. With the new notation $l(T)$ and $b(t)$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_{l(T)}^{\mathbf{CUP}}-\mathcal{R}_{T}^{\mathbf{CUP}}\right] &= \mathbb{E}\left[\left(\sum_{t=1}^{l(T)}C_t^{\mathbf{CUP}}-\sum_{t=1}^{l(T)}C_t^{\pi(S^*)}\right)-\left(\sum_{t=1}^{T}C_t^{\mathbf{CUP}}-\sum_{t=1}^{T}C_t^{\pi(S^*)}\right)\right]\\
&=\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\mathbf{CUP}}-\sum_{t=1}^{b(T)}C_t^{\mathbf{CUP}}\right]-\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\pi(S^*)}-\sum_{t=1}^{b(T)}C_t^{\pi(S^*)}\right]\\
&\quad-\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\mathbf{CUP}}-\sum_{t=1}^{l(T)}C_t^{\mathbf{CUP}}\right]+\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\pi(S^*)}-\sum_{t=1}^{l(T)}C_t^{\pi(S^*)}\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{b(T)}C_t^{\mathbf{CUP}}-\sum_{t=1}^{T}C_t^{\mathbf{CUP}}\right]-\mathbb{E}\left[\sum_{t=1}^{b(T)}C_t^{\pi(S^*)}-\sum_{t=1}^{T}C_t^{\pi(S^*)}\right] \qquad \text{(EC.25)}\\
&\geq-\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\mathbf{CUP}}-\sum_{t=1}^{l(T)}C_t^{\mathbf{CUP}}\right]+\mathbb{E}\left[\sum_{t=1}^{b(l(T))}C_t^{\pi(S^*)}-\sum_{t=1}^{l(T)}C_t^{\pi(S^*)}\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{b(T)}C_t^{\mathbf{CUP}}-\sum_{t=1}^{T}C_t^{\mathbf{CUP}}\right]-\mathbb{E}\left[\sum_{t=1}^{b(T)}C_t^{\pi(S^*)}-\sum_{t=1}^{T}C_t^{\pi(S^*)}\right], \qquad \text{(EC.26)}
\end{aligned}$$

where the inequality follows from that the sum of the first two terms of (EC.25) is non-negative since S^* minimizes the total cost between events A by Lemma EC.1.

Since the expected one-period cost difference between any two feasible policies is bounded above by $(\bar{S}-\underline{S})\max(h+\theta,p)$, and the expected number of periods between $b(l(T))$ and $l(T)$ is $1/\underline{\mu}$ which is also the same as that between $b(T)$ and T. By (EC.26), we have

$$\mathbb{E}\left[\mathcal{R}_{l(T)}^{\mathbf{CUP}}-\mathcal{R}_{T}^{\mathbf{CUP}}\right]\geq-\frac{2}{\underline{\mu}}(\bar{S}-\underline{S})\max(h+\theta,p). \tag{EC.27}$$

Thus, in what follows we shall focus on the evaluation of $\mathbb{E}[\mathcal{R}_{l(T)}^{\mathbf{CUP}}]$, the expected regret of a T-cycle problem. Since under our CUP algorithm, τ_k is a stopping time determined by demand process and previous base-stock levels and in particular, S_{k-1}. To emphasize its dependency on S_{k-1}, in the following we shall also write it as $\tau_k(S_{k-1})$, $k=1,2,\ldots$.

To derive the regret for the strongly convex case, similar as in §5, for an arbitrary S, we define $G(S,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)$ as the total cost of base-stock policy S during a fixed cycle between periods $\tau_k(S_{k-1})$ and $\tau_{k+1}(S_k)-1$, with brand new (after ordering) inventory level S in period $\tau_k(S_{k-1})$. It is important to note that not only the cost in each period is random, the number of periods in the cycle is also random. Then, the conditional expected cost of the k-th cycle is $\mathbb{E}[G(S_k,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]$, where $F_k:=((D_1,\ldots,D_{\tau_k-1});(S_1,\ldots,S_k);\tau_k)$. We shall compare this conditional expected cost with that of the bridging problem ROI, i.e., $\mathbb{E}[G(S^*,(\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]$, that starts in period $\tau_k(S_{k-1})$ with all brand new inventories. Our idea is to evaluate, for a fixed S_k, the cost difference between policies S and S^*, i.e.,

$$\mathbb{E}[G(S,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]-\mathbb{E}[G(S^*,(\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k], \tag{EC.28}$$

where the expectation is taken with respect to future random demand $(D_t;t\geq\tau_k)$.

We first show that fixing S_k, $\mathbb{E}[G(S,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]$ is strongly convex in S. By Wald's Theorem, the expected total holding and shortage cost during the cycle is

$$\frac{h\mathbb{E}[(S-D)^+]+b\mathbb{E}[(D-S)^+]}{\mathbb{P}(D\geq S_k)}.$$

By Theorem 1, the expected outdating cost during a cycle is also convex in S, hence

$$\big(\mathbb{E}[G(S,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]\big)''_S\geq\frac{(h+b)f(S)}{\mathbb{P}(D\geq S_k)}\geq\frac{(h+b)f(S)}{\mathbb{P}(D\geq\underline{S})}\geq\frac{(h+b)\lambda}{\bar{\mu}}\geq(h+b)\lambda>0. \tag{EC.29}$$

This shows that $\mathbb{E}[G(S,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]$ is strongly convex in S with parameter $(h+p)\lambda$. Therefore, applying Taylor's expansion in (EC.28) on S then set $S=S_k$ we obtain

$$\begin{aligned}&\mathbb{E}[G(S_k,\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]-\mathbb{E}[G(S^*,(\tau_k(S_{k-1}),\tau_{k+1}(S_k);\omega)\mid F_k]\\&\leq\nabla_1\mathbb{E}[G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k](S_k-S^*)-\frac{1}{2}(h+p)\lambda(S_k-S^*)^2,\end{aligned} \tag{EC.30}$$

where $\nabla_1\mathbb{E}[G(S,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k]$ is the partial subderivative of $\mathbb{E}[G(S,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k]$ with respect to the first argument S. Using (EC.30), we have

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}^{\mathbf{CUP}}_{l(T)}\right]\leq\mathbb{E}\left[\sum_{k=1}^{T}\Big(G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)-G(S^*,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\Big)\right]\\&=\mathbb{E}\left[\sum_{k=1}^{T}\Big(\mathbb{E}[G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k]-\mathbb{E}[G(S^*,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k]\Big)\right]\\&\leq\mathbb{E}\left[\sum_{k=1}^{T}\nabla_1\mathbb{E}[G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k](S_k-S^*)\right]\end{aligned}$$

$$-\mathbb{E}\left[\sum_{k=1}^{T}\frac{1}{2}(h+p)\lambda(S_k-S^*)^2\right]. \tag{EC.31}$$

Following the same argument used in (EC.18), we have, for $k=1,\dots,T$,

$$\begin{aligned}(S_{k+1}-S^*)^2 \le (S_k-S^*)^2 &- 2\eta_k(S_k-S^*)\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\\ &+\eta_k^2\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2.\end{aligned}$$

Conditioning on F_k and taking expectation with respect to future demand, yield

$$\begin{aligned}&\nabla_1\mathbb{E}[G_k(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\mid F_k](S_k-S^*) \qquad \text{(EC.32)}\\ &\le \frac{1}{2\eta_k}\Big(\mathbb{E}[(S_k-S^*)^2\mid F_k]-\mathbb{E}[(S_{k+1}-S^*)^2\mid F_k]\Big)+\frac{\eta_k}{2}\mathbb{E}\left[\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2\mid F_k\right].\end{aligned}$$

Combining (EC.31) and (EC.32), we have

$$\begin{aligned}\mathbb{E}\big[\mathcal{R}^{\mathbf{CUP}}_{l(T)}\big] &\le \mathbb{E}\Big[\sum_{k=1}^{T}\frac{1}{2\eta_k}\Big(\mathbb{E}[(S_k-S^*)^2\mid F_k]-\mathbb{E}[(S_{k+1}-S^*)^2\mid F_k]\Big)\\ &\qquad +\frac{\eta_k}{2}\mathbb{E}\big[\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2\mid F_k\big]-\frac{(h+p)\lambda}{2}\sum_{k=1}^{T}(S_k-S^*)^2\Big]\\ &=\mathbb{E}\Big[\sum_{k=1}^{T}\Big(\frac{1}{2\eta_k}[(S_k-S^*)^2-(S_{k+1}-S^*)^2\Big)\\ &\qquad +\frac{\eta_k}{2}\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2-\frac{(h+p)\lambda}{2}\sum_{k=1}^{T}(S_k-S^*)^2\Big]\\ &\le \mathbb{E}\Big[\sum_{k=1}^{T}\frac{1}{2\lambda(h+p)k}\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2-T(S_{T+1}-S^*)^2\Big]\\ &\le \mathbb{E}\Big[\sum_{k=1}^{T}\frac{1}{2\lambda(h+p)k}\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2\Big]\\ &\le \big(\max(h+\theta,p)\big)^2\cdot\frac{2-\underline{\mu}}{2(h+p)\lambda\underline{\mu}^2}\cdot\sum_{k=1}^{T}\frac{1}{k},\end{aligned} \tag{EC.33}$$

where the second inequality follows from plugging in the step-size $\eta_k=\frac{1}{(h+p)\lambda k}$, and the last inequality holds because, using the identical argument used in (EC.21), we have that for each $k=1,\dots,T$,

$$\mathbb{E}\left[\big(\nabla_1 G(S_k,(\tau_k(S_{k-1}),\tau_{k+1}(S_k));\omega)\big)^2\right]\le\big(\max(h+\theta,p)\big)^2\cdot\frac{2-\underline{\mu}}{\underline{\mu}^2}.$$

Consequently, combining (EC.33) and (EC.27), we have that with large enough T,

$$\mathbb{E}\left[\mathcal{R}^{\mathbf{CUP}}_{T}\right]\le\big(\max(h+\theta,p)\big)^2\cdot\frac{2-\underline{\mu}}{2(h+p)\lambda\underline{\mu}^2}\cdot\sum_{k=1}^{T}\frac{1}{k}+\frac{2}{\underline{\mu}}(\bar{S}-\underline{S})\max(h+\theta,p)\le K_2\cdot\log T,$$

(EC.34)

for some positive constant K_2. This completes the proof of Theorem 3.

If $\bar{\mu}$ is known *a priori*, then as seen from (EC.29), the strong convexity coefficient can be improved to $(h+b)\lambda/\bar{\mu}$. In this case, the step-size of the algorithm can be modified to $\eta_k = \left(\frac{\bar{\mu}}{\lambda(h+b)}\right)\frac{1}{k}$, then it follows from (EC.34) that the corresponding regret is reduced by the factor $\bar{\mu}$. **Q.E.D.**