POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

Vol. 29, No. 7, July 2020, pp. 1624–1649
ISSN 1059-1478 | EISSN 1937-5956 | 20 | 2907 | 1624
DOI 10.1111/poms.13178

Optimal Learning Algorithms for Stochastic Inventory Systems with Random Capacities

Weidong Chen, Cong Shi*
Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, USA, aschenwd@umich.edu, shicong@umich.edu

Izak Duenyas
Technology and Operations, Ross School of Business, University of Michigan, Ann Arbor, MI 48109, USA, duenyas@umich.edu

We propose the first learning algorithm for single-product, periodic-review, backlogging inventory systems with random production capacity. Different than the existing literature on this class of problems, we assume that the firm has neither prior information about the demand distribution nor the capacity distribution, and only has access to past demand and supply realizations. The supply realizations are *censored* capacity realizations in periods where the policy need not produce full capacity to reach its target inventory levels. If both the demand and capacity distributions were known at the beginning of the planning horizon, the well-known target interval policies would be optimal, and the corresponding optimal cost is referred to as the *clairvoyant optimal* cost. When such distributional information is not available *a priori* to the firm, we propose a cyclic stochastic gradient descent type of algorithm whose running average cost asymptotically converges to the clairvoyant optimal cost. We prove that the rate of convergence guarantee of our algorithm is $O(1/\sqrt{T})$, which is provably tight for this class of problems. We also conduct numerical experiments to demonstrate the effectiveness of our proposed algorithms.

Key words: inventory; random capacity; online learning algorithms; regret analysis
History: Received: December 2018; Accepted: January 2020 by Qi Annabelle Feng, after 1 revision.

1. Introduction

Capacity plays an important role in a production–inventory system (see Zipkin (2000) and Simchi-Levi et al. (2014)). The amount of capacity and the variability associated with this capacity affect the production plan as well as the amount of inventory that the firm will carry. As seen from our literature review in section 1.2, there has been a rich and growing literature on capacitated production–inventory systems, and this literature has demonstrated that capacitated systems are inherently more difficult to analyze compared to their uncapacitated counterparts, due to the fact that the capacity constraint makes future costs heavily dependent on the current decision. For instance, facing a capacity constraint, a mistake of under-ordering in one particular period may cause the system to be unable to produce up to the inventory target level over the next multiple periods.

The prior literature on capacitated inventory systems assumes that the stochastic future demand that the firm will face and the stochastic future capacity that the firm will have access to are given by exogenous random variables (or random processes), and the inventory decisions are made with full knowledge of future demand and capacity distributions. However, in most practical settings, the firm does not know the demand distribution *a priori*, and has to deduce the demand distribution based on the observed demand while it is producing and selling the product. Similarly, when the firm starts producing a new product on a manufacturing line, the firm may have very little idea of the variability associated with this capacity *a priori*. The uncertainty of capacity can be much more significant than the uncertainty in demand in some cases. For instance, Tesla originally stated that it had a line that would be able to build Model 3s at the rate of 5000 per week by the end of June 2017. However, Tesla was never able to reach this production rate at any time in 2017. In fact, during the entire fourth quarter of 2017, Tesla was only able to produce 2425 Model 3s according to Sparks (2018). Tesla was finally able to achieve the rate of 5000 produced cars the last week of the second quarter of 2018. However, even at the end of August 2018, Tesla was not able to achieve anywhere near an average 5000 Model 3s production rate per week. Even if we ignore ramp-up issues and assume that Tesla has finally (after one year's delay) achieved "stability," according to Bloomberg's estimate as of September 10, 2018, Tesla was only producing an average of 3857 Model 3s per week in September according to Randall and Halford (2018). Even though Tesla may have had more problems

than the average manufacturer, significant uncertainty over what production rate can be achieved at a factory is not at all uncommon. In fact, some analysts have questioned whether this line will ever be able to achieve a consistent production rate of 5000 Model 3s per week displaying the difficulty of estimating the true capacity of a production line.

Another salient example is Apple's launches of its iPhone over time. When the iPhone 6 was being introduced, there were a large number of articles (see, e.g., Brownlee (2014)) indicating that the radical redesign of Apple's smartphone would lead to a short supply of enough devices when it launched due to the increasing difficulty of producing the phone with the new design. In this case, Apple was producing the iPhone already for about seven years. However, the new generation product was significantly different so that the estimates that Apple had built of its lines' production rates based on the old products were no longer valid. Similarly, as Apple was about to launch its latest iPhone in October 2018, there were numerous reports about potential capacity problems. Sohail (2018) discussed how supply might be constrained at launch due to capacity problems. However, a month and a half after launch, Apple found that sales of its XS and XR models were less than predicted and had to resort to increasing what it offers for trade-in of previous generation iPhone models as an incentive to boost sales. Thus, even in year 11 of production of its product, Apple still has to deal with capacity and demand uncertainty and with each new generation, it has to rediscover its capacity and demand distributions. This is what has motivated us to develop a learning algorithm that helps the firm decide on how many units to produce, while it is learning about its demand and capacity distributions.

1.1. Main Results, Contributions, and Connections to Prior Work

We develop the first learning algorithm, called *the data-driven random capacity algorithm* (DRC for short), for finding the optimal policy in a periodic-review production–inventory system with random capacities, where the firm neither knows the demand distribution nor the capacity distribution *a priori*. Note that our learning algorithm is *nonparametric* in the sense that we do not assume any parametric forms of these distributions. The performance measure is the standard notion of *regret* in online learning algorithms (see Shalev-Shwartz (2012)), which is defined as the difference between the cost of the proposed learning algorithm and the *clairvoyant optimal* cost, where the clairvoyant optimal cost corresponds to the hypothetical case where the firm knew the demand and capacity distributions *a priori* and applied the optimal (target interval) policy.

Our main result is to show that the cumulative T-period regret of the DRC algorithm is bounded by $O(\sqrt{T})$, which is also theoretically the best possible for this class of problems. Our proposed learning algorithm is connected to Huh and Rusmevichientong (2009) that studied the classical multi-period stochastic inventory model and Shi et al. (2016) that considered the multi-product setting under a warehouse capacity constraint. We point out that both prior studies hinged on the myopic optimality of the clairvoyant optimal policy, that is, it suffices to examine a single-period cost function. However, the random production capacity (on how much can be produced) considered in this work is fundamentally different than the warehouse capacity (on how much can be stored) considered in Shi et al. (2016), and our problem does not enjoy myopic optimality. It is well-known in the literature that models with random production capacities are challenging to analyze, in that the current decisions will impact the cost over an extended period of time (rather than a single period). For example, an under-ordering in one particular period may cause the system to be unable to produce up to the inventory target level over the next multiple periods. Thus, we need to carefully re-examine the random capacitated problem with demand and capacity learning.

There are three main innovations in the design and analysis of our learning algorithm.

1. First, we propose a cyclic updating idea. In our setting, the "right" cycle is the so-called *production cycle*, first proposed in Ciarallo et al. (1994) to establish the *extended myopic optimality* for the random capacitated inventory systems. The production cycle is defined as the interval between successive periods in which the policy is able to attain a given base-stock level, in which one can show that the cumulative cost within a production cycle is convex in the base-stock level. Naturally, our DRC algorithm updates base-stock levels in each production cycle. Note that these production cycles (seen as renewal processes) are not *a priori* fixed but are sequentially triggered as demand and supply are realized over time. Technically, we develop explicit upper bounds on moments of the production cycle length and the associated stochastic gradient. A major challenge in the algorithm design is that the algorithm needs to determine if the current production cycle (with respect to the clairvoyant optimal system) ends before making the decision in the current period. We design for each possible scenario to gather sufficient information to determine if the target level should be updated.

2. Second, the observed capacity realizations are, in fact, censored. That is, when the plant is able to complete production (i.e., the capacity was sufficient in the current period to bring inventory up to the desired level), the actual capacity will not be revealed. This creates major challenges in the design and analysis of learning algorithms. For example, suppose that at the beginning of a period, the firm decides to produce 100 units. If the production facility has a random capacity of 80 with $\frac{1}{3}$ probability, 120 with $\frac{1}{3}$ probability, and 150 with $\frac{1}{3}$ probability, then upon producing 100 units, the firm can only confirm that the capacity in this period is not 80, but cannot decide between 120 and 150. Therefore, the firm needs to carry out *active explorations*, which is to overproduce when necessary, in order to learn the capacity correctly. If the firm employs no active explorations and believes what it observes, the firm will have an erroneous assumption on the capacity, leading to a spiral down effect.
3. Third, facing random capacity constraints, the firm may not be able to achieve the desired target inventory level as prescribed by the algorithm, and hence we keep track of a virtual (infeasible) bridging system by "temporarily ignoring" the random capacity constraints, which is used to update our target level in the next iteration. The gradient information of this virtual system needs to be correctly obtained from the demand and the censored capacity observed in the real implemented system when the random capacity constraints are imposed. Also, due to positive inventory carry-over and capacity constraints, we need to ensure that the amount of overage and underage inventory (relative to the desired target level) is appropriately bounded, to achieve the desired rate of convergence of regret.

1.2. Relevant Literature

Our work is closely related to two streams of literature: (i) capacitated stochastic inventory systems and (ii) learning algorithms for stochastic inventory systems.

Capacitated stochastic inventory systems. There has been a substantial body of literature on capacitated stochastic inventory systems. The dominant paradigm in most of the existing literature has been to formulate stochastic inventory control problems using a dynamic programming framework. This approach is effective in characterizing the structure of optimal policies. We first list the papers that consider fixed capacity. Federgruen and Zipkin (1986a, b) showed that a modified base-stock policy is optimal under both the average and discounted cost criteria. Tayur (1992), Kapuscinski and Tayur (1998), and Aviv and Federgruen (1997) derived the optimal policy under independent cyclical demands. Özer and Wei (2004) showed the optimality of modified base-stock policies in capacitated models with advance demand information. Even for these classical capacitated systems with non-perishable products, the simple structure of their optimal control policies does not lead to efficient algorithms for computing the optimal control parameters. Tayur (1992) used the shortfall distribution and the theory of storage processes to study the optimal policy for the case of i.i.d. demands. Roundy and Muckstadt (2000) showed how to obtain approximate base-stock levels by approximating the distribution of the shortfall process. Kapuscinski and Tayur (1998) proposed a simulation-based technique using infinitesimal perturbation analysis to compute the optimal policy for capacitated systems with independent cyclical demands. Özer and Wei (2004) used dynamic programming to solve capacitated models with advance demand information when the problem size is small. Levi et al. (2008) gave a 2-approximation algorithm for this class of problems. Angelus and Zhu (2017) identified the structure of optimal policies for capacitated serial inventory systems. All the papers above assume that the firm knows the stochastic demand distribution and the deterministic capacity level.

There has also been a growing body of literature on stochastic inventory systems where both demand and capacity are uncertain. When capacity is uncertain, several papers (e.g., Henig and Gerchak (1990), Federgruen and Yang (2011), Huh and Nagarajan (2010)) assumed that the firm has uncertain yield (i.e., if they start producing a certain number of products, an uncertain proportion of what they started will become finished goods). An alternative approach by Ciarallo et al. (1994) and Duenyas et al. (1997) assumed that what the firm can produce in a given time interval (e, g., a week) is stochastic (due to, e.g., unexpected downtime, unexpected supply shortage, unexpected absenteeism, etc.) and proved the optimality of extended myopic policies for uncertain capacity and stochastic demand under discounted optimal costs scenario. Güllü (1998) established a procedure to compute the optimal base stock level for uncertain capacity production–inventory systems. Wang and Gerchak (1996) extended the analysis to systems with both random capacity and random yield. Feng (2010) addressed a joint pricing and inventory control problem with random capacity and shows that the optimal policy is characterized by two critical values: a reorder point and a target safety stock. More recently, Chen et al. (2018) developed a unified transformation technique which converts a non-convex minimization problem to an equivalent convex minimization problem, and such a transformation can be used to prove the preservation of structural properties for inventory

control problems with random capacity. Feng and Shanthikumar (2018) introduced a powerful notion termed stochastic linearity in mid-point, and transformed several supply chain problems with nonlinear supply and demand functions into analytically tractable convex problems. All the papers above assume that the firm knows the stochastic demand distribution and the stochastic capacity distribution.

Learning algorithms for stochastic inventory systems. There has been a recent and growing interest in situations where the distribution of demand is not known *a priori*. Many prior studies have adopted parametric approaches (see, e.g., Lariviere and Porteus (1999), Chen and Plambeck (2008), Liyanage and Shanthikumar (2005), Chu et al. (2008)), and we refer interested readers to Huh and Rusmevichientong (2009) for a detailed discussion on the differences between parametric and nonparametric approaches.

For nonparametric approaches, Burnetas and Smith (2000) considered a repeated newsvendor problem, where they developed an algorithm that converges to the optimal ordering and pricing policy but did not give a convergence rate result. 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. Zhang et al. (2018) proposed an algorithm for the perishable inventory system. Huh et al. (2009) and Zhang et al. (2019) and Agrawal and Jia (2019) developed learning algorithms for the lost-sales inventory system with positive lead times. Yuan et al. (2019) and Ban (2019) considered fixed costs. Chen et al. (2019a, b) proposed algorithms for the joint pricing and inventory control problem with backorders and lost-sales, respectively. Chen and Shi (2020) focused on learning the best Tailored Base-Surge (TBS) policies in dual-sourcing inventory systems. 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. None of the papers surveyed above modeled random capacity with a priori unknown distribution, and we therefore need to develop new learning approaches to address this issue.

1.3. Organization and General Notation

The remainder of the study is organized as follows. In section 2, we formally describe the capacitated inventory control problem for random capacity. In section 3, we show that a target interval policy is optimal for capacitated inventory control problem with salvaging decisions. In section 4, we introduce the data-driven algorithm for random capacity under unknown demand and capacity distribution. In section 5, we carry out an asymptotic regret analysis, and show that the average T-period expected cost of our policy differs from the optimal expected cost by at most $O(\sqrt{T})$. In section 6, we compare our policy performance to the performance of two straw heuristic policies and show that simple heuristic policies used in practice may not work very well. In section 7, we conclude our study and point out plausible future research avenues.

Throughout the study, we often distinguish between a random variable and its realizations using capital and lower-case letters, respectively. For any real numbers $a, b \in \mathbb{R}$, $a^+ = \max\{a, 0\}$, $a^- = -\min\{a, 0\}$; the join operator $a \vee b = \max\{a,b\}$, and the meet operator $a \wedge b = \min\{a, b\}$.

2. Stochastic Inventory Control with Uncertain Capacity

We consider an infinite horizon periodic-review stochastic inventory planning problem with production capacity constraint. We use (time-generic) random variable D to denote random demand, and U to denote random production capacity. The random production capacity may be caused by maintenance or downtime in the production line, lack of materials, among others (see Zipkin (2000), Simchi-Levi et al. (2014), Snyder and Shen (2011)). The demand and the capacity have distribution functions $F_D(\cdot)$ and $F_U(\cdot)$, respectively, and density functions $f_D(\cdot)$ and $f_U(\cdot)$, respectively.

At the beginning of our planning horizon, the firm does not know the underlying distributions of D and U. In each period $t = 1,2,\ldots$, the sequence of events are as follows:

1. At the beginning of each period t, the firm observes the starting inventory level x_t before production. (We assume without loss of generality that the system starts empty, i.e., $x_1 = 0$.) The firm also observes the past demand and (censored) capacity realizations up to period $t-1$.
2. Then the firm decides the target inventory level s_t. If $s_t \geq x_t$, then it will try to produce $q_t = s_t - x_t$ to bring its inventory level up to s_t. Here, q_t is the target production quantity which may not be achieved due to capacity. During the period, the firm will realize its random production capacity u_t, and therefore, its

final inventory level will be $s_t \wedge (x_t + u_t)$. We *emphasize* here that the firm will not observe the actual capacity realization u_t if they meet their inventory target s_t. Thus, the firm actually observes the censored capacity $\tilde{u}_t$, that is, when the production plan cannot be fulfilled at period t, $\tilde{u}_t = u_t$; otherwise, $\tilde{u}_t = (s_t - x_t)^+ \wedge u_t$. In contrast, if $s_t < x_t$, then the firm will salvage $-q_t = x_t - s_t$ units. Notice that in our model, we allow for negative q_t, which represents salvaging. We denote the inventory level after production or salvaging as $y_t = s_t \wedge (x_t + u_t)$. If the firm decides to bring its inventory level up, it incurs a production cost $c(y_t - x_t)^+$ and if it decides to bring its inventory level down, it receives a salvage value $\theta(x_t - y_t)^+$, where c is the per-unit production cost and θ is the per-unit salvage value. We assume that $\theta \le c$.

3. At the end of the period t, after production is completed, the demand D_t is realized, and we denote its realization by d_t, which is satisfied to the maximum extent using on-hand inventory. Unsatisfied demands are *backlogged*, which means that the firm can observe full demand realization d_t in period t. The state transition can be written as $x_{t+1} = s_t \wedge (x_t + u_t) - d_t = y_t - d_t$. The overage and underage costs at the end of period t is $h(y_t - d_t)^+ + b(d_t - y_t)^+$, where h is the per unit holding cost and b is the per unit backlogging cost.

Following the system dynamics described above, we write the single-period cost as a function of s_t and x_t as follows.

$$\begin{aligned}\Omega(x_t, s_t) &= c(s_t \wedge (x_t + U_t) - x_t)^+ - \theta(x_t - s_t \wedge (x_t + U_t))^+ \\ &\quad + h(s_t \wedge (x_t + U_t) - D_t)^+ \\ &\quad + b(D_t - s_t \wedge (x_t + U_t))^+ \\ &= c(y_t - x_t)^+ - \theta(x_t - y_t)^+ + h(y_t - D_t)^+ \\ &\quad + b(D_t - y_t)^+.\end{aligned}$$

Let f_t denote the cumulative information collected up to the beginning of period t, which includes all the realized demands d, observed (censored) capacities u, and past ordering decisions s up to period $t-1$. A feasible *closed-loop* control policy π is a sequence of functions $s_t = \pi_t(x_t, f_t), t = 1, 2, \ldots$, mapping the beginning inventory x_t and f_t into the ending inventory decision s_t. The objective is to find an efficient and effective adaptive inventory control policy π, or a sequence of inventory targets $\{s_t\}_{t=1}^{\infty}$, which minimizes the long-run average expected cost

$$\limsup_{T\to\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \Omega(x_t, s_t)\right]. \tag{1}$$

Table 1 Summary of Major Notation

Symbol	Type	Description
c	Parameter	Production cost.
θ	Parameter	Salvage cost.
h	Parameter	Per unit holding cost.
b	Parameter	Per unit backlogging cost.
D_t, d_t	Parameter	Random demand and its realization in period t.
F_D, f_D	Parameter	Demand probability and density function.
U_t, u_t	Parameter	Random production capacity and its realization in period t.
F_U, f_U	Parameter	Capacity probability and density function.
s_l^* or s^*	State	Clairvoyant target product-up-to level after ordering.
s_u^*	State	Clairvoyant target salvage-down-to level after salvaging.
x_t	State	Beginning inventory level in period t.
y_t	State	Ending inventory level in period t.
s_t	Control	Target inventory level after ordering/salvaging in period t.
q_t	Control	Ordering/salvaging quantity in period t.

If there is a discount factor $\alpha \in (0,1)$, the objective becomes the total discounted cost, that is,

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^t \cdot \Omega(x_t, s_t)\right]. \tag{2}$$

The major notation used in this paper is summarized in Table 1.

3. Clairvoyant Optimal Policy (with Salvage Decisions)

To facilitate the design of a learning algorithm, we first study the clairvoyant scenario by assuming that the distributions of demand and production capacity were given *a priori*. Furthermore, we assume that the actual production capacity in each period is observed by the firm, that is, there is no capacity censoring in this clairvoyant case. The clairvoyant case is useful as it serves as a lower bound on the cost achievable by the learning model. For the case where the firm can only raise its inventory (without any salvage decisions), Ciarallo et al. (1994) showed that a *produce-up-to* policy is optimal. A *minor contribution* of this study is to extend their policy by enabling the firm to salvage extra goods with salvage price θ at the beginning of each period before the demand is realized. The firm incurs production cost c per-unit good if it decides to produce and receives a salvage value of θ (i.e., incurring a salvage cost $-\theta$) per-unit good if it decides to salvage, and $c \le \theta$.

We shall describe a *target interval policy*, and show that it is optimal. A target interval policy is characterized by two threshold values (s_l^*, s_u^*) such that if the starting inventory level $x < s_l^*$, we order up to s_l^*, if $x > s_u^*$, we salvage down to s_u^*, and if $s_l^* \le x \le s_u^*$, we do nothing. Note that target interval policy has been introduced in a number of earlier papers. In fact, the

structure of this policy was first identified by Eberly and Van Mieghem (1997) and the term target interval policy was first used by Angelus and Porteus (2002).

ASSUMPTION 1. *We make the following assumptions on the demand and capacity distributions.*

1. *The demands* $D_1, \ldots, D_T$ *and the capacities* $U_1, \ldots, U_T$ *are independently and identically distributed* (*i.i.d.*) *continuous random variables, respectively. Also, the demand* D_t *and the capacity* U_t *are independent across all time periods* $t \in \{1,\ldots T\}$.
2. *The* (*time generic*) *demand and capacity* D *and* U *have a bounded support* $[0, \bar{d}]$ *and a bounded support* $[0, \bar{u}]$, *respectively. We also assume that* $\mathbb{E}[U] > \mathbb{E}[D]$ *to ensure the system stability.*
3. *The* (*clairvoyant*) *optimal produce-up-to level* s_l^* *lies in a bounded interval* $[0, \bar{s}]$, *that is,* $s_l^* \in [0, \bar{s}]$.

Assumption 1(a) assumes the stationarity of the underlying production–inventory system to be jointly learned and optimized over time. Assumption 1(b) ensures the stability of the system, that is, the system can clear all the backorders from time to time. Assumption 1(c) assumes that the firm knows an upper bound (potentially a loose one) on the optimal ordering levels. These assumptions are mild and standard in inventory learning literature (see, e.g., Huh and Rusmevichientong (2009), Huh et al. (2009), Zhang et al. (2019, 2018)). We also remark here that an important future research direction is to incorporate non-stationarity of the demand and capacity processes, which would require a significant methodological breakthrough.

3.1. Optimal Policy for the Single Period Problem with Salvaging Decisions

We first use a single-period problem to illustrate the idea of target interval policy, and then extend it to the multi-period problem with salvage decisions.

PROPOSITION 1. *For the single period problem, a target interval policy is optimal. More specifically, there exist two threshold levels* s_l^* *and* s_u^* *such that the optimal policy can be described as follows:*

1. *When* $s_l^* \le x \le s_u^*$, *the firm decides to do nothing.*
2. *When* $x < s_l^*$, *the firm decides to produce to bring inventory up to* s_l^* *as close as possible.*
3. *When* $s_u^* < x$, *the firm decides to salvage and bring inventory down to* s_u^*.

The three situations discussed above can be readily illustrated in Figure 1. The two curves are labeled "$q \ge 0$" and "$q < 0$," respectively. The solid curve is the effective cost function $\Omega(y)$, which consists of curve "$q \ge 0$" for $s \ge x$, and curve "$q < 0$" for $s < x$.

3.2. Optimal Policy for the Multi-Period Problem with Salvaging Decisions

Next, we derive the optimal policy for the multi-period problem with salvaging decisions.

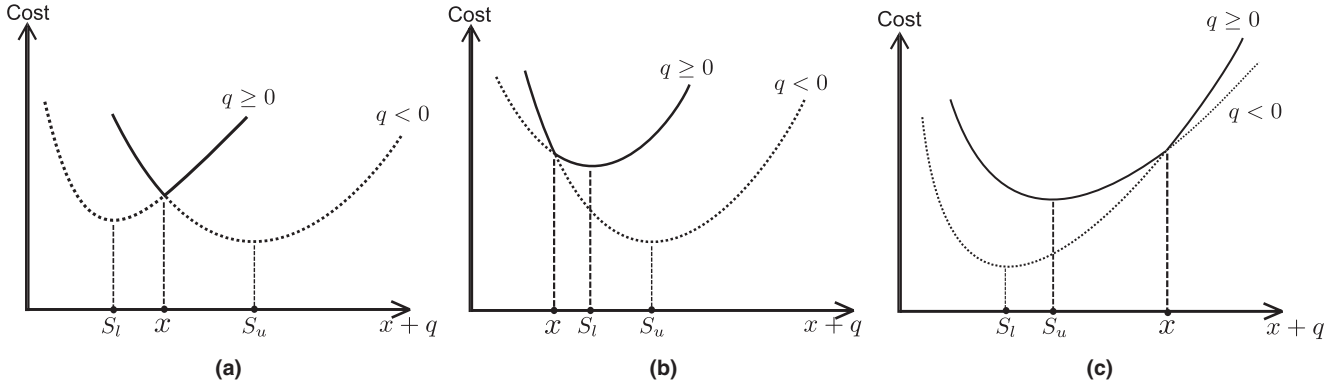
PROPOSITION 2.

1. *For the T-period finite-horizon problem with salvaging decisions, a target interval policy is optimal. More specifically, for each period* $t = 1,\ldots, T$, *there exist two time-dependent threshold levels* $s_{t,l}^*$ *and* $s_{t,u}^*$ *such that the optimal target level* s_t^* *satisfies*

$$s_t^* = \begin{cases} s_{t,l}^*, & x_t < s_{t,l}^*, \\ x_t, & s_{t,l}^* \le x_t \le s_{t,u}^*, \\ s_{t,u}^*, & x_t > s_{t,u}^*. \end{cases}$$

2. *For both the infinite horizon discounted problem* (2) *with salvaging decisions and the average cost problem* (1) *with salvaging decisions, a target interval policy is optimal. More specifically, there exist two time-invariant threshold levels* s_l^* *and* s_u^* *such that the optimal target level* s_t^* *satisfies*

Figure 1 Illustration of a Target Interval Policy



$$s_t^* = \begin{cases} s_l^*, & x_t < s_l^*, \\ x_t, & s_l^* \le x_t \le s_u^*, \\ s_u^*, & x_t > s_u^*. \end{cases}$$

Note that for the finite time horizon case, the optimal target level depends on a pair of *time-dependent* threshold levels, whereas for the infinite horizon case, the optimal interval policy depends on a pair of *time-invariant* threshold levels. Since the clairvoyant benchmark is chosen with respect to the infinite horizon problem, our goal is to find the optimal target interval (s_l^*, s_u^*). We have shown that if the firm has the option to salvage extra goods at the beginning of each period, then it will choose to salvage extra goods if the starting inventory is high enough. In the full-information problem, we can immediately conclude that in the infinite horizon problem, the salvage decision will only be made in the first period when the initial starting inventory is higher than s_u^*. This is because after salvaging down to s_u^* in the first period, the inventory level will gradually be consumed down below s_l^* and after that, the inventory level will never exceed s_l^* again, due to the stationary demand assumption. Thus, the optimal produce-up-to level s_l^* is the same as the optimal produce-up-to level, denoted by s^*, in Ciarallo et al. (1994) without salvaging options, and an extended myopic policy described therein is also optimal for the infinite horizon average cost setting. In the remainder of this study, we will use s_l^* and s^* interchangeably. However, we must *emphasize* here that in the learning version of the problem, since we do not know the demand and capacity distributions (and of course s_l^* or s^*), we need to *actively explore* the inventory space, and salvaging decisions will be made in our online learning algorithm (more frequently in the beginning phase).

4. Nonparametric Learning Algorithms

As discussed in section 1, in many practical scenarios, the firm neither knows the distribution of demand D nor the distribution of production capacity U at the beginning of the planning horizon. Instead, the firm has to rely on the observable demand and capacity realizations over time to make adaptive production decisions. More precisely, in each period t, the firm can observe the realized demand d_t as well as the observed production capacity $\tilde{u}_t$. In our model, while d_t is the true demand realization (since the demands are backlogged), the observed production capacity $\tilde{u}_t$ is, in fact, *censored*. More explicitly, the censored capacity $\tilde{u}_t = (s_t - x_t)^+ \wedge u_t$. That is, suppose the firm wants to raise the starting inventory level x_t to some target level s_t. If the true realized production capacity $u_t > (s_t - x_t)^+$, then the firm cannot observe the uncensored capacity realization u_t. Our objective is to find an efficient and effective learning production control policy whose long-run average cost converges to the clairvoyant optimal cost (had the distributional information of both the random demand and the random capacity been given *a priori*) at a provably tight convergence rate.

4.1. The Notion of Production Cycles

It is well-known in the literature that the optimal policy for a capacitated inventory system cannot be solved myopically, that is, the control that minimizes a single-period cost is not optimal. Moreover, when capacities are random, the per-period cost function is non-convex, due to the fact that the decision is truncated by a random variable (see Chen et al. (2018) and Feng and Shanthikumar (2018)). Thus, one cannot run the stochastic gradient descent algorithms period by period. To overcome this difficulty, we partition the set of time periods into carefully designed learning cycles, and update our production target levels from cycle to cycle, instead of from period to period.

We now formally define these learning cycles. Given that we produce up to the target level s_t in some period t and then use the same target level s_t for all subsequent periods, we define a *production cycle* as the set of successive periods starting from period t until the next period in which we are able to produce up to s_t again. Mathematically, let τ_j denote the starting period of the j^{th} production cycle. Then, for any given initial target level $s_1 \in [0, \bar{s}]$, we have

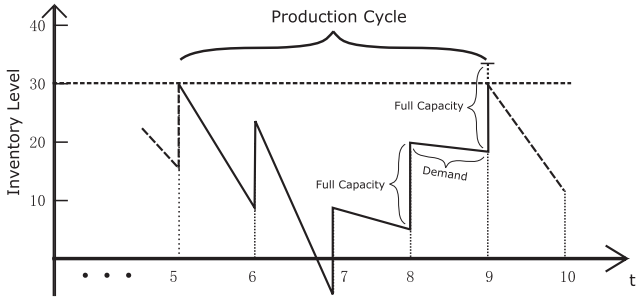
$$\tau_1 = 1,$$
$$\tau_j = \min\left\{ t \ge \tau_{j-1} + 1 \;\middle|\; x_t + u_t \ge s_{\tau_{j-1}} \right\}, \text{ for all } j \ge 2.$$

For convenience, we call s_{τ_j} the *cycle target level* for production cycle j. We let l_j be the cycle length of the j^{th} production cycle, that is, $l_j = \tau_{j+1} - \tau_j$.

Figure 2 gives a simple graphical example of a production cycle. Suppose the target production level $s_5 = 30$ and the realized capacity levels $u_t = 15$ for $t = 5,\ldots,9$. In periods 6,7,8, we are not able to attain the target level s_5 even if we produce the full capacity in these periods, whereas we are able to do so in period 9. Therefore, this production cycle runs from period 5 to period 9. Note that in period 9, we could only observe the censored capacity $\tilde{u}_9 = 11$ (instead of the true realized capacity $u_9 = 15$), because we only need to produce 11 to attain the target level.

The definition of these production cycles is motivated by the idea of *extended myopic policies*, which we shall discuss next. In the full-information (clairvoyant) case with stationary demand, the structural results in section 3 imply that if the system starts with initial inventory s^* (for simplicity we drop the subscript from the optimal produce-up-to level s_l^*), then the optimal policy is a modified base-stock policy, that is, in each period t,

Figure 2 An Illustration of a Production Cycle



$$y_t = \begin{cases} s^*, & \text{if} x_t + u_t \geq s^*, \\ x_t + u_t, & \text{if} x_t + u_t < s^*. \end{cases}$$

In this case, our definition of production cycles reduces to

$$\tau_1 = 1,$$
$$\tau_j = \min\left\{ t \geq \tau_{j-1} + 1 \;\middle|\; y_t = s^* \right\}, \text{ for all } j \geq 2.$$

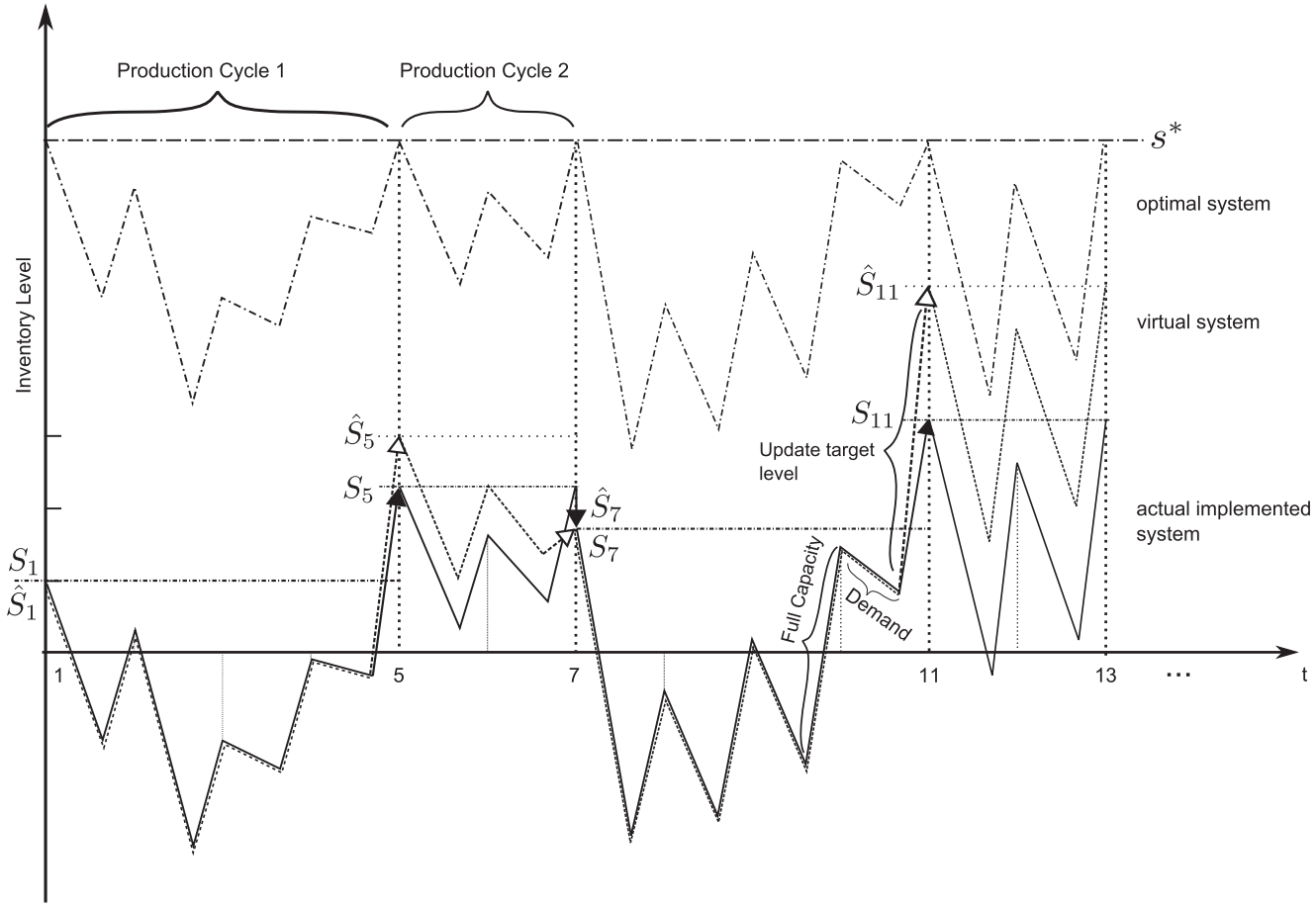
In other words, the optimal system forms a sequence of production cycles whose cycle target levels are all set to be s^*, which is also illustrated at the top portion of Figure 3. Ciarallo et al. (1994) showed that the extended myopic policy, which is obtained by merely minimizing the expected total cost within a single production cycle, is optimal. (They also provided a computationally tractable procedure to compute this s^* with known demand and capacity distributions.)

The above discussion has motivated us to design a nonparametric learning algorithm that updates the modified base-stock levels in a cyclic way, in which the sequence of production cycle costs in our system will eventually converge to the production cycle cost of the optimal system. We *emphasize* again that the (clairvoyant) optimal system does not need to salvage since s^* is known, whereas our system needs to actively explore the inventory space to learn the value of s^* and thus salvaging can happen frequently in the beginning phase of the learning algorithm.

4.2. The Data-Driven Random Capacity Algorithm

With the definition of production cycles, we shall describe our data-driven random capacity algorithm (DRC for short). The DRC algorithm keeps track of two systems in parallel, and also ensures that both systems share the same production cycles as in the

Figure 3 An Illustration of the Algorithmic Design



optimal system (which uses the same optimal base-stock level s^* in every period). The optimal system is depicted using dash-dot lines shown at the top of Figure 3. The optimal system starts at optimal base-stock level s^*, and uses s^* as target level in every period.

The first system that the DRC algorithm keeps track of is a virtual (or ideal) system, which starts from an arbitrary inventory level $\hat{s}_1$. The DRC algorithm maintains a triplet $(\hat{s}_t, \hat{y}_t, \hat{x}_t)$ in each period t, where $\hat{s}_t$ is the virtual target level, $\hat{y}_t$ is the virtual inventory level, and $\hat{x}_t$ is the virtual starting inventory level. At the beginning of each production cycle j, namely, in period τ_j, the DRC algorithm computes the (desired) virtual cycle target level $\hat{s}_{\tau_j}$, and *artificially adjusts* the virtual inventory level $\hat{y}_{\tau_j} = \hat{s}_{\tau_j}$ by temporarily ignoring the random capacity constraint in that period. For all subsequent periods $t \in [\tau_j + 1, \tau_{j+1} - 1]$ within production cycle j, the DRC algorithm sets the virtual target production level $\hat{s}_t = \hat{s}_{\tau_j}$ and runs the virtual system as usual (facing the same demands and random capacity constraints as in the actual implemented system), that is, $\hat{y}_t = \hat{s}_t \wedge (\hat{x}_t + u_t)$ and $\hat{x}_{t+1} = \hat{y}_t - d_t$. Figure 3 gives an example of the evolution of a virtual system, as depicted using dotted lines.

The second system is the actual implemented system, which starts from an arbitrary inventory level $s_1 = \hat{s}_1$. The DRC algorithm maintains a triplet (s_t, y_t, x_t) in each period t, where s_t is the target production level, y_t is the actual attained inventory level, and x_t is the actual starting inventory level. Different than the virtual system described above, at the beginning of each production cycle j, namely, in period τ_j, the DRC algorithm tries to reach the (desired) virtual target level $\hat{s}_{\tau_j}$ but may fail to do so due to random capacity constraints. The resulting inventory level y_{τ_j} may possibly be lower than $\hat{s}_{\tau_j}$. Nevertheless, to keep the production cycle synchronized with that of the optimal system, we simply set the cycle target level $s_{\tau_j} = y_{\tau_j}$, and keep the target production level the same within the production cycle, that is, $s_t = s_{\tau_j}$ for all $t \in [\tau_j, \tau_{j+1} - 1]$. Figure 3 gives an example of the evolution of an actual implemented system (as depicted using solid lines).

We now present the detailed description of the DRC algorithm.

The Data-Driven Random Capacity Algorithm (DRC)

Step 0. (*Initialization.*) In the first period $t = 1$, set the initial inventory $x_1 \in [0, \bar{s}]$ arbitrarily. We set both the target level and the virtual target level the same as the initial inventory, that is, $s_1 = \hat{s}_1 = x_1$. Then we also have the actual attained inventory level $y_1 = x_1$ and the virtual inventory level $\hat{y}_1 = \hat{x}_1 = x_1$. Initialize the counter for production cycles $j = 1$, and set $t = \tau_1 = 1$.

Step 1. (*Updating the Virtual System.*) The algorithm updates the virtual target level in period $t + 1$ by

$$\hat{s}_{t+1} = \begin{cases} \mathbf{Proj}_{[0,\bar{s}]}\left(\hat{s}_{\tau_j} - \eta_j \cdot \sum_{k=\tau_j}^{t} \mathcal{G}_k(\hat{s}_{\tau_j})\right), & \text{if } t = \tau_j, \\ \hat{s}_{\tau_j}, & \text{if } t > \tau_j, \end{cases}$$

where $\mathcal{G}_k(\hat{s}_{\tau_j})$

$$= \begin{cases} h, & \text{if } \hat{s}_{\tau_j} \wedge (\hat{x}_k + u_k) \geq d_k, \\ -b, & \text{otherwise.} \end{cases}$$

Note that the projection operator $\mathbf{Proj}_{[0,\bar{s}]}(x) = \max\{0, \min\{x, \bar{s}\}\}$. The step-size is chosen to be

$$\eta_j = \frac{\gamma}{\sqrt{\sum_{k=1}^{j} l_k}}, \quad \text{where} \quad l_k = \tau_{k+1} - \tau_k,$$

where $\gamma > 0$ is a constant (to be optimized later for the tightest theoretical regret bound).

The evolution of the virtual system is given as follows,

$$\hat{y}_t = \begin{cases} \hat{s}_{\tau_j} - \sum_{i=\tau_j}^{t-1} d_i + \sum_{i=\tau_j+1}^{t} u_i, & \text{for } t > \tau_j, \\ \hat{s}_{\tau_j}, & \text{for } t = \tau_j, \end{cases} \quad \text{and}$$

$$\hat{x}_{t+1} = \hat{y}_t - d_t.$$

Step 2. (*Updating the Actual Implemented System.*) We have the following cases when updating the actual implemented system based on $\hat{s}_t$.

1. If $\hat{s}_{t+1} \geq s_{\tau_j}$, then we try to produce up to $\hat{s}_{t+1}$, and the actual inventory level y_{t+1} will be

$$y_{t+1} = \begin{cases} \hat{s}_{t+1}, & \text{if } x_{t+1} + u_{t+1} \geq \hat{s}_{t+1}, \\ x_{t+1} + u_{t+1}, & \text{if } x_{t+1} + u_{t+1} < \hat{s}_{t+1}. \end{cases}$$

   (i). If $s_{\tau_j} \leq y_{t+1} \leq \hat{s}_{t+1}$, we start a new production cycle $j + 1$, by setting the starting period of this new cycle $\tau_{j+1} = t + 1$. Correspondingly, we set the virtual cycle target level $\hat{s}_{\tau_{j+1}} = \hat{s}_{t+1}$, and the actual implemented cycle target level $s_{\tau_{j+1}} = y_{t+1}$. We then increase the value of j by one.

   (ii). If $y_{t+1} < s_{\tau_j}$, we are still in the same production cycle j, and thus we set $s_{t+1} = s_{\tau_j}$.

2. If $\hat{s}_{t+1} < s_{\tau_j}$, then we first try to produce up to s_{τ_j} (instead of $\hat{s}_{t+1}$) , and the actual inventory level y_{t+1} will be

$$y_{t+1} = \begin{cases} s_{\tau_j}, & \text{if } x_{t+1} + u_{t+1} \geq s_{\tau_j}, \\ x_{t+1} + u_{t+1}, & \text{if } x_{t+1} + u_{t+1} < s_{\tau_j}. \end{cases}$$

   (i). If $y_{t+1} = s_{\tau_j}$, we salvage our inventory level down to $y_{t+1} = \hat{s}_{t+1}$. We then start a new

production cycle j+1, by setting the starting period of this new cycle $\tau_{j+1} = t + 1$. Correspondingly, we set the virtual cycle target level $\hat{s}_{\tau_{j+1}} = \hat{s}_{t+1}$, and the actual implemented cycle target level $s_{\tau_{j+1}} = \hat{s}_{t+1}$. We then increase the value of j by one.

(ii) If $y_{t+1} < s_{\tau_j}$, we are still in the same production cycle j, and thus we set $s_{t+1} = s_{\tau_j}$.

We then increase the value of t by one, and go to Step 1. If $t = T$, terminate the algorithm.

4.3. Overview of the Data-Driven Random Capacity Algorithm

In Step 1, we update the virtual system using the online stochastic gradient descent method. In each period t of any given cycle j, the DRC algorithm tries to minimize the total expected cost associated with production cycle j by updating the virtual target level using a gradient estimator $\sum_{k=\tau_j}^{t} \mathcal{G}_k(\hat{s}_{\tau_j})$ of the total cost accrued from period τ_j to period t. We shall show in Lemma 4 below that $G_j(\hat{s}_{\tau_j}) = \sum_{k=\tau_j}^{\tau_{j+1}-1} \mathcal{G}_k(\hat{s}_{\tau_j})$ is the sample-path cycle cost gradient of production cycle j. Note that $G_j(\hat{s}_{\tau_j})$ is the sample-path cycle cost gradient for the *virtual system*. However, we could only observe the demand and censored capacity information in the *actual implemented system*, and the key question is whether this information is sufficient to evaluate this $G_j(\hat{s}_{\tau_j})$ correctly.

LEMMA 1. *The sample-path cycle cost gradient of the virtual system* $G_j(\hat{s}_{\tau_j}) = \sum_{k=\tau_j}^{\tau_{j+1}-1} \mathcal{G}_k(\hat{s}_{\tau_j})$ *for every cycle* $j \geq 1$ *can be evaluated correctly by only using the observed demand and censored capacity information of the actual implemented system.*

PROOF OF LEMMA 1. It suffices to show that for each period $k = \tau_j, \ldots, \tau_{j+1} - 1$, the cost gradient estimator $\mathcal{G}_k(\hat{s}_{\tau_j})$ can be evaluated correctly. We have the following two cases.

1. If $k = \tau_j$, that is, the production cycle j starts in period k, we must have $x_k + \tilde{u}_k \geq s_{\tau_{j-1}}$ by our definition of production cycle. In addition, we observe the full capacity $\tilde{u}_i = u_i$ in period $i = \tau_{j-1} + 1, \ldots, k-1$ but only observe the censored capacity $\tilde{u}_k \leq u_k$ in period k.

   (i) if $s_k = \hat{s}_k$, by the system dynamics we have

$$\hat{s}_k = s_k = x_k + \tilde{u}_k \leq \hat{x}_k + \tilde{u}_k \leq \hat{x}_k + u_k,$$

where the first inequality holds because by our algorithm design, we always have $s_{\tau_{j-1}} \leq \hat{s}_{\tau_{j-1}}$ for all $j = 2,3,\ldots$, and then

$$x_k = s_{\tau_{j-1}} - \sum_{i=\tau_{j-1}}^{\tau_j - 1} d_i + \sum_{i=\tau_{j-1}+1}^{\tau_j - 1} u_i \leq \hat{s}_{\tau_{j-1}} - \sum_{i=\tau_{j-1}}^{\tau_j - 1} d_i + \sum_{i=\tau_{j-1}+1}^{\tau_j - 1} u_i = \hat{x}_k.$$

Hence, the event $\{\hat{s}_{\tau_j} \wedge (\hat{x}_k + u_k) \geq d_k\}$ is equivalent to $\{\hat{s}_{\tau_j} \geq d_k\}$, and therefore, we can evaluate $\mathcal{G}_k(\hat{s}_{\tau_j})$ correctly.

   (ii). if $s_k < \hat{s}_k$, we have produced full capacity and therefore observe the full capacity $\tilde{u}_k = u_k$. Then the event $\{\hat{s}_{\tau_j} \wedge (\hat{x}_k + u_k) \geq d_k\}$ is equivalent to $\{\hat{s}_{\tau_j} \wedge (\hat{x}_k + \tilde{u}_k) \geq d_k\}$, and therefore we can evaluate $\mathcal{G}_k(\hat{s}_{\tau_j})$ correctly.

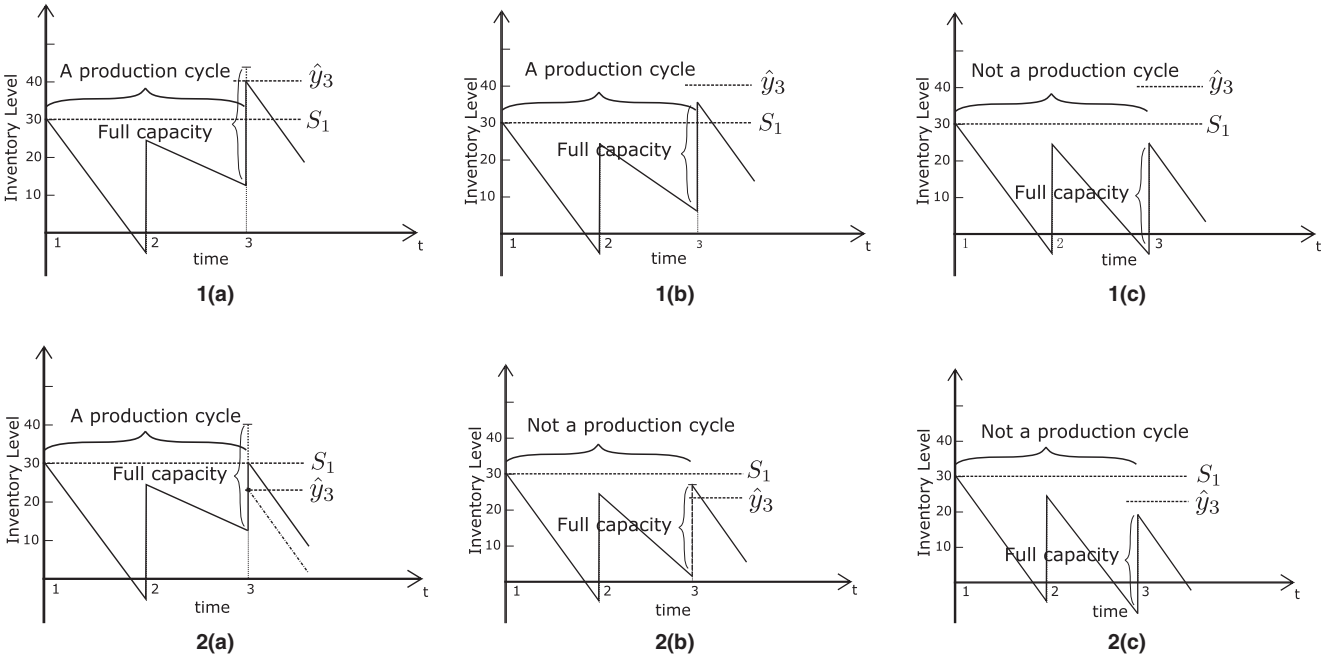
2. In contrast, if $k \in [\tau_j + 1, \tau_{j+1} - 1]$, that is, then we are still in the current production cycle j. In this case, we always produce at full capacity, and therefore, we observe the full capacity $\tilde{u}_k = u_k$. Then the event $\{\hat{s}_{\tau_j} \wedge (\hat{x}_k + u_k) \geq d_k\}$ is equivalent to $\{\hat{s}_{\tau_j} \wedge (\hat{x}_k + \tilde{u}_k) \geq d_k\}$, and therefore, we can evaluate $\mathcal{G}_k(\hat{s}_{\tau_j})$ correctly. □

Combining the above two cases yields the desired the result.

In Step 2, we compare $\hat{s}_{t+1}$ and s_{τ_j} to decide how to update the actual implemented system. We have two cases. The first case is when $\hat{s}_{t+1} \geq s_{\tau_j}$. We want to produce up to the new target level $\hat{s}_{t+1}$ instead of s_{τ_j}. If the actual implemented inventory level $y_{t+1} \geq s_{\tau_j}$, we know that the current production cycle ends because we have achieved at least s_{τ_j}, and then we shall start the next production cycle. In order to perfectly align the production cycle with that of the optimal system when $\hat{s}_{t+1} \geq y_{t+1} \geq s_{\tau_j}$, we should set the next cycle target level $s_{\tau_{j+1}} = y_{t+1}$. Otherwise, we produce at full capacity, and stay in the same production cycle, which is also synchronized with the optimal production cycle. The second case is when $\hat{s}_{t+1} < s_{\tau_j}$. We first produce up to the current cycle target level s_{τ_j} to check whether we can start the next production cycle. If s_{τ_j} is achieved, we shall start the next production cycle and salvage the inventory level down to $y_{t+1} = \hat{s}_{t+1}$ and also set the new cycle target level $s_{\tau_{j+1}} = \hat{s}_{t+1}$. Otherwise, we produce at full capacity, and stay in the same production cycle, which is also synchronized with the optimal production cycle.

The central idea here is to *align* the production cycles of the actual implemented system (as well as the virtual bridging system) with those of the (clairvoyant) optimal system, even while updating our cycle target level at the beginning of each production cycle. As illustrated in Figure 3, the optimal system knows s^* *a priori* and keeps using the

Figure 4 A Schematic Illustration of All Possible Scenarios



target level s^* (i.e., the optimal modified base-stock level) in every period t. Whenever the target level s^* is achieved, we start the next production cycle. However, in the learning problem, the firm does not know s^* and needs to constantly update the cycle target level at the beginning of each production cycle. Due to the discrepancy between the new and the previous target levels, it is crucial to design an algorithm that can determine whether the current production cycle ends, and whether we should adopt the new target level in the very same period. Figure 4 shows the possible scenarios. The scenarios 1(a), 1(b), and 1(c) show the case when $\hat{s}_{t+1} \geq s_{\tau_j}$. In this case, we always raise the inventory to $\hat{s}_{t+1}$ as much as possible. If $\hat{s}_{t+1}$ is achieved, we know that the production cycle ends. Even if $\hat{s}_{t+1}$ is not achieved, we know that we produce at full capacity and then can readily determine whether the production cycle ends (by checking if we reach at least s_{τ_j}). The scenarios 2(a), 2(b), and 2(c) show the case when $\hat{s}_{t+1} < s_{\tau_j}$. In this case, we always raise the inventory to s_{τ_j} as much as possible to determine whether the production cycle ends (by checking if we reach exactly s_{τ_j}). We salvage the inventory level down to $\hat{s}_{t+1}$ only if the production cycle ends. Note that *active explorations* are needed in the sense that sometimes the learning algorithm will have to produce up and then salvage down in the same period, so as to obtain unbiased capacity information. Technically, doing so ensures that the production cycles are perfectly aligned between the actual implemented system and the clairvoyant optimal system.

4.4. Discussion of the Data-Driven Random Capacity Algorithm without Censoring

We have elaborated the challenges of facing censored capacity in the previous sections. The censored capacity comes from the fact that the production is terminated once the inventory level reaches the target level, and as a result, the true capacity will not be revealed.

Now, we shall discuss the setting in which the firm has access to the uncensored capacity information. There are the following two cases: (1) If the firm knows the true capacity before making the production decision, then the firm knows if a production cycle ends in the current period. In this case, the firm only needs to update the virtual target level at the end of the production cycle. The firm will always produce up to the virtual target level, without the need of any salvaging options. This case leads to a simplified DRC algorithm. (2) In contrast, if the firm knows the true capacity only after making the production decision, then the firm does not know if a production cycle ends in the current period. In this case, the firm still requires the use of the full-fledged DRC algorithm (as designed for the setting with censored capacity information).

5. Performance Analysis of the Data-Driven Random Capacity Algorithm

We carry out a performance analysis of our proposed DRC algorithm. The performance measure is the natural notion of regret, which is defined as the difference between the cost incurred by our nonparametric learning algorithm DRC and the clairvoyant optimal

cost (where the demand and production capacity distribution are both known *a priori*). That is, for any $T \geq 1$,

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(\Omega(x_t, s_t) - \Omega(x_t, s^*))\right],$$

where s_t is the target level prescribed by the DRC algorithm for period t, and s^* is the clairvoyant optimal target level. We note that our clairvoyant benchmark is chosen with respect to the infinite horizon problem, and the regret quantifies the cumulative loss of running our learning algorithm for any $T \geq 1$ periods, compared to this stationary benchmark.

Theorem 1 below states the main result of this study.

THEOREM 1. *For stochastic inventory systems with demand and capacity learning, the cumulative regret $\mathcal{R}_T$ of the data-driven random capacity algorithm (DRC) is upper bounded by $O(\sqrt{T})$. In other words, the average regret $\mathcal{R}_T/T$ approaches to 0 at the rate of $O(1/\sqrt{T})$.*

REMARK 1. Let $\mu = \mathbb{E}[U] - \mathbb{E}[D]$, the difference between expected capacity and expected demand. We define $\upsilon = 2\mu^2/(\bar{u} + \bar{d})^2$ and $X_1 = (h \vee b)l_1 - \sum_{t=\tau_1+1}^{\tau_2} U_t + \sum_{t=\tau_1}^{\tau_2-1} D_t$, and then further define $\alpha = -\mathbb{E}[X_1]$ and $\sigma^2 = Var[X_1]$ and $\beta = \mathbb{E}[X_1^3]$. The optimal constant γ in the step size (that gives rise to the tightest theoretical regret bound) is given by

$$\gamma = \frac{\bar{s}}{\sqrt{\begin{array}{c}(h \vee b)^2\left(\frac{1}{\upsilon} + \frac{2}{\upsilon^2} + \frac{2}{\upsilon^3}\right) + 2(h \vee b)^2 \frac{\bar{s}}{\mu}\frac{\sigma}{\alpha} e^{\frac{6\beta}{\sigma^3} + \frac{\alpha}{\sigma}} \\ +2(c + \theta)(h \vee b)\frac{\sigma}{\alpha} e^{\frac{6\beta}{\sigma^3} + \frac{\alpha}{\sigma}}\end{array}}},$$

and the associated constant K in the regret bound of Theorem 1 is given by

$$K = \bar{s}\sqrt{\begin{array}{c}(h \vee b)^2\left(\frac{1}{\upsilon} + \frac{2}{\upsilon^2} + \frac{2}{\upsilon^3}\right) + 2(h \vee b)^2 \frac{\bar{s}}{\mu}\frac{\sigma}{\alpha} e^{\frac{6\beta}{\sigma^3} + \frac{\alpha}{\sigma}} \\ +2(c + \theta)(h \vee b)\frac{\sigma}{\alpha} e^{\frac{6\beta}{\sigma^3} + \frac{\alpha}{\sigma}}\end{array}}.$$

The proposed DRC algorithm is the first learning algorithm for random capacitated inventory systems, which achieves a square-root regret rate. Moreover, this square-root regret rate is *unimprovable*, even for the repeated newsvendor problem without inventory carryover and with infinite capacity, which is a special case of our problem.

PROPOSITION 3. *Even in the case of uncensored demand, the square-root regret rate is tight.*

PROOF OF PROPOSITION 3. The proof follows Proposition 1 in Zhang et al. (2019) for the repeated newsvendor problem (without inventory carryover and with infinite capacity). □

The remainder of this study is to establish the regret upper bound in Theorem 1. For each $j \geq 1$, if we adopt the cycle target level s_{τ_j} and also artificially set the initial inventory level $x_{\tau_j} = s_{\tau_j}$, we can then express the cost associated with the production cycle j as

$$\begin{aligned}\Theta(s_{\tau_j}) &= \sum_{t=\tau_j+1}^{\tau_{j+1}} c\Big(s_{\tau_j} \wedge (x_t + U_t) - x_t\Big)^+ \\ &\quad + \sum_{t=\tau_j}^{\tau_{j+1}-1}\Big[h\Big(s_{\tau_j} \wedge (x_t + U_t) - D_t\Big)^+ \\ &\quad + b\Big(D_t - s_{\tau_j} \wedge (x_t + U_t)\Big)^+\Big] \\ &= \sum_{t=\tau_j+1}^{\tau_{j+1}-1} cU_t + c(s_{\tau_j} - x_{\tau_{j+1}}) \\ &\quad + \sum_{t=\tau_j}^{\tau_{j+1}-1}\Big[h\Big(s_{\tau_j} \wedge (x_t + U_t) - D_t\Big)^+ \\ &\quad + b\Big(D_t - s_{\tau_j} \wedge (x_t + U_t)\Big)^+\Big] \\ &= \sum_{t=\tau_j+1}^{\tau_{j+1}-1} cU_t + c\left(\sum_{t=\tau_j}^{\tau_{j+1}-1} D_t - \sum_{t=\tau_j+1}^{\tau_{j+1}-1} U_t\right) \\ &\quad + \sum_{t=\tau_j}^{\tau_{j+1}-1}\Big[h\Big(s_{\tau_j} \wedge (x_t + U_t) - D_t\Big)^+ \\ &\quad + b\Big(D_t - s_{\tau_j} \wedge (x_t + U_t)\Big)^+\Big],\end{aligned} \tag{3}$$

where the second equality comes from the fact that we always produce at full capacity within a production cycle, except for the last period in which we are able to reach the target level. The third equality follows from expressing

$$\begin{aligned}x_{\tau_{j+1}} &= x_{\tau_j} + \sum_{t=\tau_j+1}^{\tau_{j+1}-1} U_t - \sum_{t=\tau_j}^{\tau_{j+1}-1} D_t \\ &= s_{\tau_j} + \sum_{t=\tau_j+1}^{\tau_{j+1}-1} U_t - \sum_{t=\tau_j}^{\tau_{j+1}-1} D_t.\end{aligned}$$

Now, we use J to denote the total number of production 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 $\tau_{J+1} - 1 = T$, that is, $s_{\tau_{J+1}} = s_{\tau_J}$.) By the construction of the DRC algorithm, we can write the cumulative regret as

$$\begin{aligned}\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\Omega(x_t,s_t)-\Omega(x_t,s^*)\right]\\ &= \mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j})+\sum_{j=1}^{J}\left(c\left(s_{\tau_{j+1}}-s_{\tau_j}\right)^+\right.\right.\\ &\quad\left.\left.+\theta\left(s_{\tau_j}-s_{\tau_{j+1}}\right)^+\right)-\sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(x_t,s^*)\right]\\ &= \mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j})-\sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(x_t,s^*)\right]\\ &\quad+\mathbb{E}\left[\sum_{j=1}^{J}\left(c\left(s_{\tau_{j+1}}-s_{\tau_j}\right)^+\!+\theta\left(s_{\tau_j}-s_{\tau_{j+1}}\right)^+\right)\right]\\ &= \mathbb{E}\left[\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})-\sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(x_t,s^*)\right]\\ &\quad+\mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j})-\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})\right]\\ &\quad+\mathbb{E}\left[\sum_{j=1}^{J}\left(c\left(s_{\tau_{j+1}}-s_{\tau_j}\right)^+\!+\theta\left(s_{\tau_j}-s_{\tau_{j+1}}\right)^+\right)\right],\end{aligned}$$

where on the right-hand side of the fourth equality, the first term is the production cycle cost difference between using the virtual target level $\hat{s}_{\tau_j}$ and using the clairvoyant optimal target level s^*. The second term is the production cycle cost difference between using the actual implemented target level s_{τ_j} and using the virtual target level $\hat{s}_{\tau_j}$. The third term is the cumulative production and salvaging costs incurred by adjusting the production cycle target levels.

To prove Theorem 1, it is clear that it suffices to establish the following set of results.

Proposition 4. *For any $J \geq 1$, there exists a constant $K_1 \in \mathbb{R}^+$ such that*

$$\mathbb{E}\left[\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})-\sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(x_t,s^*)\right]\leq K_1\sqrt{T}.$$

Proposition 5. *For any $J \geq 1$, there exists a constant $K_2 \in \mathbb{R}^+$ such that*

$$\mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j})-\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})\right]\leq K_2\sqrt{T}.$$

Proposition 6. *For any $J \geq 1$, there exists a constant $K_3 \in \mathbb{R}^+$ such that*

$$\mathbb{E}\left[\sum_{j=1}^{J}\left(c\left(s_{\tau_{j+1}}-s_{\tau_j}\right)^+\!+\theta\left(s_{\tau_j}-s_{\tau_{j+1}}\right)^+\right)\right]\leq K_3\sqrt{T}.$$

5.1. Several Key Building Blocks for the Proof of Theorem 1

Before proving Propositions 4– 6, we first establish some key preliminary results.

Recall that the production cycle defined in section 4.1 is the interval between successive periods in which the policy is able to attain a given base-stock level. We first show that the cumulative cost within a production cycle is convex in the base-stock level.

Lemma 2. *The production cycle cost $\Theta(s)$ is convex in s along every sample path.*

Proof of Lemma 2. It suffices to analyze the first production cycle cost (with $x_1 = s_1$)

$$\begin{aligned}\Theta(s_1) &= \sum_{t=2}^{\tau_2-1}cU_t+c\left(\sum_{t=1}^{\tau_2-1}D_t-\sum_{t=2}^{\tau_2-1}U_t\right)\\ &\quad+\sum_{t=1}^{\tau_2-1}\left[h(s_1\wedge(x_t+U_t)-D_t)^+\right.\\ &\quad\left.+b(D_t-s_1\wedge(x_t+U_t))^+\right].\end{aligned}$$

Taking the first derivative of $\Theta(s_1)$ with respect to s_1, we have

$$\Theta'(s_1)=\sum_{t=1}^{\tau_2-1}\left(h(\xi_t^+(s_1))-b(\xi_t^-(s_1))\right),\tag{4}$$

where $\xi_t^+(s_1)=\mathbb{1}\left\{s_1-\sum_{t'=1}^{t}D_{t'}+\sum_{t'=2}^{t}U_{t'}\geq 0\right\}$ and $\xi_t^-(s_1)=\mathbb{1}\left\{s_1-\sum_{t'=1}^{t}D_{t'}+\sum_{t'=2}^{t}U_{t'}<0\right\}$ are indicator functions of the positive inventory left-over and the unsatisfied demand at the end of period t, respectively.

For any given $\delta > 0$, we have

$$\Theta'(s_1+\delta)=\sum_{t=1}^{\tau_2-1}\left[h\left(\xi^+(s_1+\delta)\right)-b(\xi^-(s_1+\delta))\right].$$

It is clear that when the target level increases, the positive inventory left-over will also increase, i.e, $\xi^+(s_1+\delta)\geq\xi^+(s_1)$. Similarly, we also have $\xi^-(s_1+\delta)\leq\xi^-(s_1)$. Therefore, we have $\Theta'(s_1+\delta)\geq\Theta'(s_1)$ for any value of s_1, and thus $\Theta(\cdot)$ is convex.

Given the convexity result, our DRC algorithm updates base-stock levels in each production cycle. Note that these production cycles (as renewal processes) are not *a priori* fixed but are sequentially triggered as demand and capacity realize over time. Therefore, we need to develop an upper bound on the moments of a random production cycle. The proof of Lemma 3 relies on building an upward drifting random walk with U_t as upward step and D_t as downward step, wherein the chance of hitting a level below zero is exponentially small due to concentration inequalities. Since the ending of a production cycle corresponds to the situation where the random walk hits zero, the second moment of its length of the current production cycle can be bounded.

LEMMA 3. *The second moment of the length of a production cycle $\mathbb{E}[l_j^2]$ is bounded for all cycle j.*

PROOF OF LEMMA 3. By the definition of a production cycle in section 4.1, we have

$$\begin{aligned}&\mathbb{P}\{l_j = l\}\\ &\quad= \mathbb{P}\Bigg\{U_{\tau_j+1} - D_{\tau_j} < 0, \ldots, \sum_{t=\tau_j+1}^{\tau_j+l-1} U_t - \sum_{t=\tau_j}^{\tau_j+l-2} D_t < 0,\\ &\qquad \sum_{t=\tau_j+1}^{\tau_j+l} U_t - \sum_{t=\tau_j}^{\tau_j+l-1} D_t \geq 0\Bigg\}.\end{aligned}$$

Since D_t and U_t are both i.i.d., so is l_j. Let M_k be an upward drifting random walk, more precisely, $M_k = \sum_{t=1}^{k}(U_t - D_t)$. Then we have, by letting $\mu = \mathbb{E}[U_t - D_t]$ and $\upsilon = 2\mu^2/(\bar{u} + \bar{d})^2$,

$$\begin{aligned}\mathbb{E}\left[l_j^2\right] &= \sum_{k=1}^{\infty} k^2 \mathbb{P}(M_1 < 0, , M_{k-1} < 0, M_k \geq 0)\\ &\leq \sum_{k=1}^{\infty} k^2 \mathbb{P}(M_{k-1} - (k-1)\mu < -(k-1)\mu)\\ &\leq \sum_{k=1}^{\infty} k^2 \exp\left(-\frac{2(k-1)\mu^2}{(\bar{u}+\bar{d})^2}\right)\\ &\leq \int_0^{\infty} (k+1)^2 \exp\left(-\frac{2k\mu^2}{(\bar{u}+\bar{d})^2}\right)\\ dk &= \frac{1}{\upsilon} + \frac{2}{\upsilon^2} + \frac{2}{\upsilon^3} < \infty,\end{aligned}$$

where the second inequality follows from the Hoeffding's inequality.

We also need to develop an upper bound on the cycle cost gradient.

LEMMA 4. *For any $j \geq 1$, the function $G_j(s) = \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathcal{G}_t(s)$ is the sample-path cycle cost gradient of production cycle j, where s is the cycle target level. Moreover, $G_j(\cdot)$ has a bounded second moment, that is, $\mathbb{E}[G_j^2(s)] < \infty$ for any s.*

PROOF OF LEMMA 4. From the definition of $G_j(s)$ and (4), it is clear that

$$G_j(s) = \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathcal{G}_t(s) = \sum_{t=\tau_j}^{\tau_{j+1}-1} \left[h(\xi_t^+(s)) - b(\xi_t^-(s))\right] = \Theta'(s).$$

Moreover, we have

$$\begin{aligned}\mathbb{E}\left[G_j^2(s)\right] &= \mathbb{E}\left[\left(\sum_{t=1}^{\tau_2-1}\left(h(\xi_t^+(s_1)) - b(\xi_t^-(s_1))\right)\right)^2\right]\\ &\leq \mathbb{E}\left[(h \vee b)^2 l_j^2\right] = (h \vee b)^2 \mathbb{E}\left[l_j^2\right] < \infty,\end{aligned}$$

where the last inequality follows from Lemma 3.

5.2. Proof of Proposition 4

Proposition 4 provides an upper bound on the production cycle cost difference between using the virtual target level $\hat{s}_{\tau_j}$ and using the clairvoyant optimal target level s^*. The proof follows a similar argument used in the general stochastic approximation literature (see Nemirovski et al. (2009)) as well as the online convex optimization literature (see Hazan (2016)). The main point of departure is due to the *a priori* random cycles, and therefore the proof relies crucially on Lemmas 3 and 4 previously established.

By optimality of s^*, we have $\mathbb{E}[\Omega(s^*, s^*)] = \inf_x\{\mathbb{E}[\Omega(x, s^*)]\}$, that is, s^* minimizes the expected single period cost. Also notice that the length of a production cycle is independent of the cycle target level being implemented. Thus, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j}) - \sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(x_t, s^*)\right]\\ &\leq \mathbb{E}\left[\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j}) - \sum_{j=1}^{J}\sum_{t=\tau_j}^{\tau_{j+1}-1}\Omega(s^*, s^*)\right] \qquad (5)\\ &= \mathbb{E}\left[\sum_{j=1}^{J}\left(\Theta(\hat{s}_{\tau_j}) - \Theta(s^*)\right)\right].\end{aligned}$$

By the sample path convexity of $\Theta(\cdot)$ shown in Lemma 2, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{J}\left(\Theta(\hat{s}_{\tau_j}) - \Theta(s^*)\right)\right] &\leq \sum_{j=1}^{J}\mathbb{E}\left[\nabla\Theta(\hat{s}_{\tau_j})(\hat{s}_{\tau_j} - s^*)\right]\\ &= \sum_{j=1}^{J}\mathbb{E}\left[G_j(\hat{s}_{\tau_j})(\hat{s}_{\tau_j} - s^*)\right]. \qquad (6)\end{aligned}$$

By the definition of $\hat{s}_{\tau_{j+1}}$ in the DRC algorithm,

$$\begin{aligned}\mathbb{E}\left(\hat{s}_{\tau_{j+1}}-s^*\right)^2 &\le \mathbb{E}\left(\hat{s}_{\tau_j}-\eta_j G_j(\hat{s}_{\tau_j})-s^*\right)^2\\ &=\mathbb{E}\left(\hat{s}_{\tau_j}-s^*\right)^2+\mathbb{E}\left(\eta_j G_j(\hat{s}_{\tau_j})\right)^2\\ &\quad-\mathbb{E}\left[2\eta_j G_j(\hat{s}_{\tau_j})(\hat{s}_{\tau_j}-s^*)\right]\\ &=\mathbb{E}\left(\hat{s}_{\tau_j}-s^*\right)^2+\mathbb{E}[\eta_j]\mathbb{E}\left(G_j(\hat{s}_{\tau_j})\right)^2\\ &\quad-2\mathbb{E}[\eta_j]\mathbb{E}\left[G_j(\hat{s}_{\tau_j})(\hat{s}_{\tau_j}-s^*)\right],\end{aligned}$$

where the second equality holds because the step-size η_j is independent of $\hat{s}_{\tau_j}$ and $G_j(\hat{s}_{\tau_j})$. Thus,

$$\mathbb{E}\left[G_j(\hat{s}_{\tau_j})(\hat{s}_{\tau_j}-s^*)\right]\le\frac{1}{2\mathbb{E}[\eta_j]}\left(\mathbb{E}\left(\hat{s}_{\tau_j}-s^*\right)^2-\mathbb{E}\left(\hat{s}_{\tau_{j+1}}-s^*\right)^2\right)+\frac{1}{2}\mathbb{E}\left[\eta_j\left(G_j(\hat{s}_{\tau_j})\right)^2\right]. \tag{6}$$

Combining (6) and (7), we have

$$\begin{aligned}&\sum_{j=1}^{J}\mathbb{E}\left[\nabla\Theta(\hat{s}_{\tau_j})(\hat{s}_{\tau_j}-s^*)\right]\\ &\le\sum_{j=1}^{J}\left(\frac{1}{2\mathbb{E}[\eta_j]}\left(\mathbb{E}\left(\hat{s}_{\tau_j}-s^*\right)^2-\mathbb{E}\left(\hat{s}_{\tau_{j+1}}-s^*\right)^2\right)+\frac{1}{2}\mathbb{E}\left[\eta_j\left(G_j(\hat{s}_{\tau_j})\right)^2\right]\right)\\ &=\frac{1}{2\mathbb{E}[\eta_1]}\mathbb{E}(\hat{s}_{\tau_1}-s^*)^2-\frac{1}{2\mathbb{E}[\eta_j]}\mathbb{E}\left(\hat{s}_{\tau_{j+1}}-s^*\right)^2+\frac{1}{2}\sum_{j=2}^{J}\left(\frac{1}{\mathbb{E}[\eta_j]}-\frac{1}{\mathbb{E}[\eta_{j-1}]}\right)\mathbb{E}\left(\hat{s}_{\tau_j}-s^*\right)^2+\sum_{j=1}^{J}\frac{\mathbb{E}\left[\eta_j\left(G_j(\hat{s}_{\tau_j})\right)^2\right]}{2}\\ &\le 2\bar{s}^2\left(\frac{1}{2\mathbb{E}[\eta_1]}+\frac{1}{2}\sum_{j=2}^{J}\left(\frac{1}{\mathbb{E}[\eta_j]}-\frac{1}{\mathbb{E}[\eta_{j-1}]}\right)\right)+\frac{\mathbb{E}[\left(G_j(\hat{s}_{\tau_j})\right)^2]}{2}\sum_{j=1}^{J}\mathbb{E}[\eta_j]\\ &=\frac{\bar{s}^2}{\mathbb{E}[\eta_J]}+\frac{\mathbb{E}[\left(G_j(\hat{s}_{\tau_j})\right)^2]}{2}\sum_{j=1}^{J}\mathbb{E}[\eta_j]K_1\sqrt{T},\end{aligned}$$

where the last inequality holds due to Lemma 4 (the bounded second moment of $G(\cdot)$) and

$$\sum_{j=1}^{J}\mathbb{E}[\eta_j]=\gamma\sum_{j=1}^{J}\mathbb{E}\left[1/\sqrt{\sum_{i=1}^{j}l_i}\right]\le\gamma\sum_{t=1}^{T}1/\sqrt{t}\le 2\gamma\sqrt{T}.$$

5.3. Proof of Proposition 5

Proposition 5 provides an upper bound on the production cycle cost difference between using the actual implemented target level s_{τ_j} and using the virtual target level $\hat{s}_{\tau_j}$. The main idea of this proof on a high level is to set up an upper bounding stochastic process that resembles the waiting time process of a **GI/GI/1** queue. A similar argument appeared Huh and Rusmevichientong (2009) and Shi et al. (2016). There are two differences. First, the mapping to the waiting time process is more involved in the presence of random capacities. In the above two papers, the resulting level is always higher than the target level, whereas the resulting level could be either higher or lower than the target level in our setting. Second, this study needs to bound the difference in cycle target levels (relying on Lemmas 3 and 4), rather than per-period target levels.

By the definition of production cycle cost (3), we have

$$\begin{aligned}\mathbb{E}\left[\Theta(s_{\tau_j})-\Theta(\hat{s}_{\tau_j})\right]&=\mathbb{E}\left[\sum_{t=\tau_j}^{\tau_{j+1}-1}\left[h\left(s_{\tau_j}\wedge(x_t+U_t)-D_t\right)^+ +b\left(D_t-s_{\tau_j}\wedge(x_t+U_t)\right)^+\right]\right.\\ &\quad\left.-\sum_{t=\tau_j}^{\tau_{j+1}-1}\left[h\left(\hat{s}_{\tau_j}\wedge(x_t+U_t)-D_t\right)^+ +b\left(D_t-\hat{s}_{\tau_j}\wedge(x_t+U_t)\right)^+\right]\right]\\ &\le\mathbb{E}\left[\sum_{t=1}^{l_j-1}(h\vee b)|s_{\tau_j}-\hat{s}_{\tau_j}|\right]\\ &\le\mathbb{E}[l_j](h\vee b)|s_{\tau_j}-\hat{s}_{\tau_j}|,\end{aligned}$$

where the second inequality holds due to the Wald's Theorem using the fact that l_j is independent of s_{τ_j} and $\hat{s}_{\tau_j}$, and the first inequality follows from the fact that for any $t\in[\tau_j,\tau_{j+1}-1]$, we have

$$\begin{aligned}&\mathbb{E}\left[\left[h\left(s_{\tau_j}\wedge(x_t+U_t)-D_t\right)^+ +b\left(D_t-s_{\tau_j}\wedge(x_t+U_t)\right)^+\right]\right.\\ &\left.-\left[h\left(\hat{s}_{\tau_j}\wedge(x_t+U_t)-D_t\right)^+ +b\left(D_t-\hat{s}_{\tau_j}\wedge(x_t+U_t)\right)^+\right]\right]\\ &\le\mathbb{E}\left[h\left(s_{\tau_j}\wedge(x_t+U_t)-\hat{s}_{\tau_j}\wedge(x_t+U_t)\right)^+ +b\left(\hat{s}_{\tau_j}\wedge(x_t+U_t)-s_{\tau_j}\wedge(x_t+U_t)\right)^+\right]\\ &\le(h\vee b)\left|s_{\tau_j}-\hat{s}_{\tau_j}\right|.\end{aligned}$$

Thus, to prove Proposition 5, it suffices to prove

$$\mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j})-\sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})\right]\le\mathbb{E}[l_j](h\vee b)\mathbb{E}\left[\sum_{j=1}^{J}|s_{\tau_j}-\hat{s}_{\tau_j}|\right]\le O(\sqrt{T}).$$

Next, we consider an auxiliary stochastic process $(Z_j \mid j\ge 0)$ defined by

$$Z_{j+1}=\left[Z_j+\frac{\gamma\lambda_j}{\sqrt{\sum_{t=1}^{j}l_t}}-v_j\right]^+, \tag{8}$$

where the random variables $\lambda_j=(h\vee b)l_j$, and $v_j=\sum_{t=\tau_j+1}^{\tau_{j+1}}U_t-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t$, and $Z_0=0$. Moreover, since we know that in period τ_{j+1}, the production cycle ends, we must have

$$v_j=\sum_{t=\tau_j+1}^{\tau_{j+1}}U_t-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t\ge 0.$$

Now we want to relate $|\hat{s}_{\tau_j}-s_{\tau_j}|$ to the stochastic process defined above. We can see from the DRC algorithm that the only situation when the virtual target level cannot be achieved is when $\hat{s}_{\tau_j}>s_{\tau_j}$. When $\hat{s}_{\tau_j}\le s_{\tau_j}$, we can salvage extra inventory and achieve the virtual target level. Therefore, we relate $|\hat{s}_{\tau_j}-s_{\tau_j}|$ with the stochastic process Z_j.

Lemma 5. *For any $j\ge 1$,*

$$\mathbb{E}\left[\sum_{j=1}^{J}|s_{\tau_j}-\hat{s}_{\tau_j}|\right]\le\mathbb{E}\left[\sum_{j=1}^{J}Z_j\right],$$

where $\{Z_j, j\ge 1\}$ is the stochastic process we define above.

Proof of Lemma 5. All the stochastic comparisons within this proof are with probability one. When $\hat{s}_{\tau_{j+1}}<x_{\tau_{j+1}}+U_{\tau_{j+1}}$, we have $\hat{s}_{\tau_{j+1}}-s_{\tau_{j+1}}=0\le Z_{j+1}$. When $\hat{s}_{\tau_{j+1}}>x_{\tau_{j+1}}+U_{\tau_{j+1}}$, we have $s_{\tau_{j+1}}=x_{\tau_{j+1}}+U_{\tau_{j+1}}=s_{\tau_j}-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t+\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t+U_{\tau_j+1}$. Therefore, we have

$$\begin{aligned}\left|\hat{s}_{\tau_{j+1}}-s_{\tau_{j+1}}\right|&=\hat{s}_{\tau_{j+1}}-s_{\tau_{j+1}}=\mathbf{Proj}_{[0,\bar{s}]}\left(\hat{s}_{\tau_j}-\eta_jG_j(\hat{s}_{\tau_j})\right)-s_{\tau_{j+1}}\le\left|\mathbf{Proj}_{[0,\bar{s}]}\left(\hat{s}_{\tau_j}-\eta_jG_j(\hat{s}_{\tau_j})\right)\right|-s_{\tau_{j+1}}\\&\le\left|\hat{s}_{\tau_j}-\eta_jG_j(\hat{s}_{\tau_j})\right|-s_{\tau_j}+\left(\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t-\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t\right)-U_{\tau_{j+1}}\\&\le\left|\hat{s}_{\tau_j}-s_{\tau_j}-\eta_jG_j(\hat{s}_{\tau_j})\right|+\left(\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t-\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t\right)-U_{\tau_{j+1}}\\&\le\left|\hat{s}_{\tau_j}-s_{\tau_j}\right|+\left|\eta_jG_j(\hat{s}_{\tau_j})\right|-\left(\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t\right)\\&\le\left|\hat{s}_{\tau_j}-s_{\tau_j}\right|+\eta_j(h\vee b)\cdot l_j-\left(\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t\right),\end{aligned}$$

where the first equality holds because following the DRC algorithm, we always have $s_{\tau_j}\le\hat{s}_{\tau_j}$. The third inequality holds because s_{τ_j} is always nonnegative. This is because the virtual target level is truncated to be nonnegative all the time, and we update the actual implemented target level when the production cycle ends, which means after the previous actual implemented target level is achieved. Since $s_1\ge 0$, $s_{\tau_j}\ge 0$ for all j. The fourth inequality holds because of the triangular inequality and the last inequality holds because $|G_j(\hat{s}_{\tau_j})|\le(h\vee b)\cdot l_j$.

Therefore, from the above claim, we have

$$\left|s_{\tau_{j+1}}-\hat{s}_{\tau_{j+1}}\right|\le\left[|s_{\tau_j}-\hat{s}_{\tau_j}|+\eta_j(h\vee b)l_j-\left(\sum_{t=\tau_j+1}^{\tau_{j+1}-1}U_t-\sum_{t=\tau_j}^{\tau_{j+1}-1}D_t\right)\right]^+.$$

Comparing to (8), we have

$$\eta_j(h\vee b)l_j\le\frac{\gamma\lambda_j}{\sqrt{\sum_{t=1}^{j}l_t}},$$

and since $s_1-\hat{s}_1=0$, it follows, from the recursive definition of Z_j, that $|s_{\tau_{j+1}}-\hat{s}_{\tau_{j+1}}|\le Z_{j+1}$ holds with probability one. Summing up both sides of the inequality completes the proof.

We observe that the stochastic process Z_j is very similar to the waiting time in a **GI/GI/1** queue, except that the service time is scaled by $\gamma/\sqrt{\sum_{i=1}^{j}l_i}$ in each production cycle j. Now consider a **GI/GI/1** queue $(W_j|j\ge 0)$ defined by the following Lindley's equation: $W_0=0$, and

$$W_{j+1}=\left[W_j+\lambda_j-v_j\right]^+, \tag{9}$$

where the sequences λ_j and v_j consist of independent and identically distributed random variables

(only dependent upon the distributions of D and U). Let $\varphi_0 = 0$, $\varphi_1 = \inf\{t \geq 1 : W_j = 0\}$ and for $t \geq 1$, $\varphi_{t+1} = \inf\{t > \varphi_t : W_j = 0\}$. Let $B_t = \varphi_t - \varphi_{t-1}$. The random variable W_j is the waiting time of the j^{th} customer in the **GI/GI/1** queue, where the inter-arrival time between the j^{th} and $j+1^{th}$ customers is distributed as v_j, and the service time is distributed as λ_j. Then, B_t is the length of the t^{th} busy period. Let $\rho = \mathbb{E}[\lambda_1]/\mathbb{E}[v_1]$ represent the system utilization. Note that if $\rho < 1$, then the queue is stable, and the random variable B_t is independent and identically distributed.

We invoke the following result from Loulou (1978) to bound $\mathbb{E}[B_t]$, the expected busy period of a **GI/GI/1** queue with inter-arrival distribution v and service time λ.

Lemma 6. (*Loulou* (1978)) *Let* $X_j = \lambda_j - v_j$, *and* $\alpha = -\mathbb{E}[X_1]$. *Let* σ^2 *be the variance of* X_1. *If* $\mathbb{E}[X_1]^3 = \beta < \infty$, *and* $\rho < 1$,

$$\mathbb{E}[B_1] \leq \frac{\sigma}{\alpha}\exp\left(\frac{6\beta^3}{\sigma^3} + \frac{\alpha}{\sigma}\right).$$

For each $n \geq 1$, let the random variable $i(n)$ denote the index t such that B_t contains n. This means that the n^{th} customer is within the $B_{i(n)}$ busy period. Since B_t is i.i.d., we know that $\mathbb{E}[B_{i(n)}] = \mathbb{E}[B_t] = \mathbb{E}[B_1]$.

Lemma 7. *For any period* $t \geq 1$, *we have*

$$\mathbb{E}\left[\sum_{j=1}^{J} Z_j\right] \leq 2\gamma(h \vee b)\mathbb{E}[B_1]\sqrt{T}.$$

Proof of Lemma 7. As defined above, the stochastic process $Z_{j+1} = \left[Z_j + \frac{\gamma\lambda_j}{\sqrt{\sum_{i=1}^{j} l_i}} - v_j\right]^+$. Since Z_j can be interpreted as the waiting time in the **GI/GI/1** queueing system, we can rewrite Z_j as

$$\begin{aligned} Z_j &= \sum_{j'=1}^{j}\left(\frac{\gamma\lambda_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}} - v_{j'}\right)\mathbb{1}\left[j' \in B_{i(j)}\right] \\ &\leq \sum_{j'=1}^{j}\frac{\gamma\lambda_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}}\mathbb{1}\left[j' \in B_{i(j)}\right]. \end{aligned} \tag{10}$$

We then bound the total waiting time of sequence Z_j by only considering the cumulative service times as follows:

$$\begin{aligned} \mathbb{E}\left[\sum_{j=1}^{J} Z_j\right] &= \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{j}\frac{\gamma\lambda_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}}\mathbb{1}[j' \in B_{i(j)}]\right] \\ &\leq \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J}\frac{\gamma(h \vee b)l_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}}\mathbb{1}[j' \in B_{i(j)}]\right] \\ &\leq \mathbb{E}\left[\sum_{j'=1}^{J}\frac{\gamma(h \vee b)l_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}}\sum_{j=1}^{J}\mathbb{1}[j' \in B_{i(j)}]\right] \\ &= \mathbb{E}\left[\sum_{j'=1}^{J}\frac{\gamma(h \vee b)l_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}}B_{i(j')}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{\gamma(h \vee b)}{\sqrt{t}}B_{i(t)}\right], \end{aligned}$$

where the last inequality holds because

$$\sum_{j'=1}^{J}\frac{l_{j'}}{\sqrt{\sum_{i=1}^{j'} l_i}} \leq \sum_{t=1}^{T}\frac{1}{\sqrt{t}}, \quad \text{where} \quad T = \sum_{j'=1}^{J} l_{j'}.$$

Thus, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{j=1}^{J} Z_j\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{\gamma(h \vee b)}{\sqrt{t}}B_{i(t)}\right] \\ &= \gamma(h \vee b)\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\right]\mathbb{E}[B_{i(t)}] \leq 2\gamma(h \vee b)\sqrt{T}\mathbb{E}[B_1], \end{aligned} \tag{11}$$

where the last inequality follows from the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T} - 1$. Combining 10 and 11 completes the proof.

Combining Lemmas 5 and 7, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{j=1}^{J}\Theta(s_{\tau_j}) - \sum_{j=1}^{J}\Theta(\hat{s}_{\tau_j})\right] &\leq \mathbb{E}\left[\sum_{j=1}^{J}\gamma(h \vee b)(\hat{s}_{\tau_j} - s_{\tau_j})\right] \\ &\leq \gamma(h \vee b)\mathbb{E}[l_1]\mathbb{E}\left[\sum_{j=1}^{J} Z_j\right] \\ &\leq 2\gamma(h \vee b)^2\mathbb{E}[l_1]\mathbb{E}[B]\sqrt{T}, \end{aligned}$$

where both $\mathbb{E}[B]$ and $\mathbb{E}[l_1]$ are bounded constants. This completes the proof for Proposition 5.

5.4. Proof of Proposition 6

Proposition 6 provides an upper bound on the cumulative production and salvaging costs incurred by adjusting the production cycle target levels.

The main idea of this proof on a high level is to use the fact that the cycle target levels of the actual

implemented system are getting closer to the ones of the virtual system over time, and each change in the cycle target level can be sufficiently bounded, resulting in an upper bound on the cumulative production and salvaging costs.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{J} c\left(s_{\tau_{j+1}} - s_{\tau_j}\right)^+\right] \le \mathbb{E}\left[\sum_{j=1}^{J} c\left(\hat{s}_{\tau_{j+1}} - s_{\tau_j}\right)^+\right] \\
=&\mathbb{E}\left[\sum_{j=1}^{J} c\left(\mathbf{Proj}_{[0,\bar{s}]}\left(\hat{s}_{\tau_j} - \eta_j \cdot G_j(\hat{s}_{\tau_j})\right) - s_{\tau_j}\right)^+\right] \\
\le&\mathbb{E}\left[\sum_{j=1}^{J} c\left(\left(\hat{s}_{\tau_j} - \eta_j \cdot G_j(\hat{s}_{\tau_j})\right) - s_{\tau_j}\right)^+\right] \\
\le&\mathbb{E}\left[\sum_{j=1}^{J} c\left|\hat{s}_{\tau_j} - s_{\tau_j}\right| + \sum_{j=1}^{J} c\left|\eta_j \cdot G_j(\hat{s}_{\tau_j})\right|\right] \le K_4\sqrt{T},
\end{aligned}$$

where K_4 is some positive constant. The result trivially holds if $s_{\tau_{j+1}} \le s_{\tau_j}$. Now, consider the case where $s_{\tau_{j+1}} > s_{\tau_j}$, that is, the firm produces. The first inequality holds because if the firm produces, we must have $s_{\tau_{j+1}} \le \hat{s}_{\tau_{j+1}}$ by the construction of DRC. The second inequality holds because $s_{\tau_j} \ge 0$. The third inequality holds by the triangular inequality. The last inequality is due to the fact that $\sum_{j=1}^{J}\left|\hat{s}_{\tau_j} - s_{\tau_j}\right| \le O(\sqrt{T})$ from Proposition 5, and

$$\sum_{j=1}^{J} c\left|\eta_j \cdot G_j(\hat{s}_{\tau_j})\right| \le c\gamma(h \vee b)\sum_{j=1}^{J}\frac{l_j}{\sqrt{\sum_{i=1}^{j} l_i}} \le 2c\gamma(h \vee b)\sqrt{T}.$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{J} \theta(s_{\tau_j} - s_{\tau_{j+1}})^+\right] = \mathbb{E}\left[\sum_{j=1}^{J} \theta(s_{\tau_j} - \hat{s}_{\tau_{j+1}})^+\right] \\
=&\mathbb{E}\left[\sum_{j=1}^{J} \theta\left(s_{\tau_j} - \mathbf{Proj}_{[0,\bar{s}]}\left(\hat{s}_{\tau_j} - \eta_j \cdot G_j(\hat{s}_{\tau_j})\right)\right)^+\right] \\
\le&\mathbb{E}\left[\sum_{j=1}^{J} \theta\left(s_{\tau_j} - \left(\hat{s}_{\tau_j} - \eta_j \cdot G_j(\hat{s}_{\tau_j})\right)\right)^+\right] \\
\le&\mathbb{E}\left[\sum_{j=1}^{J} \theta\left|\hat{s}_{\tau_j} - s_{\tau_j}\right| + \sum_{j=1}^{J} \theta\left|\eta_j \cdot G_j(\hat{s}_{\tau_j})\right|\right] \le K_5\sqrt{T},
\end{aligned}$$

where K_5 is some positive constant. The result trivially holds if $s_{\tau_j} \le s_{\tau_{j+1}}$. Now, consider the case where $s_{\tau_j} > s_{\tau_{j+1}}$, that is, the firm salvages. The first equality holds because if the firm salvages, we must have $s_{\tau_{j+1}} = \hat{s}_{\tau_{j+1}}$ by the construction of DRC. The first inequality holds because $\bar{s} \ge s_{\tau_j}$. The second inequality holds by the triangular inequality. The last inequality follows the same idea as in the first part of this section.

Combing the above two parts completes the proof of Proposition 6.

Finally, Theorem 1 is a direct consequence of Propositions 4–6, which gives us the desired regret upper bound.

6. Numerical Experiments

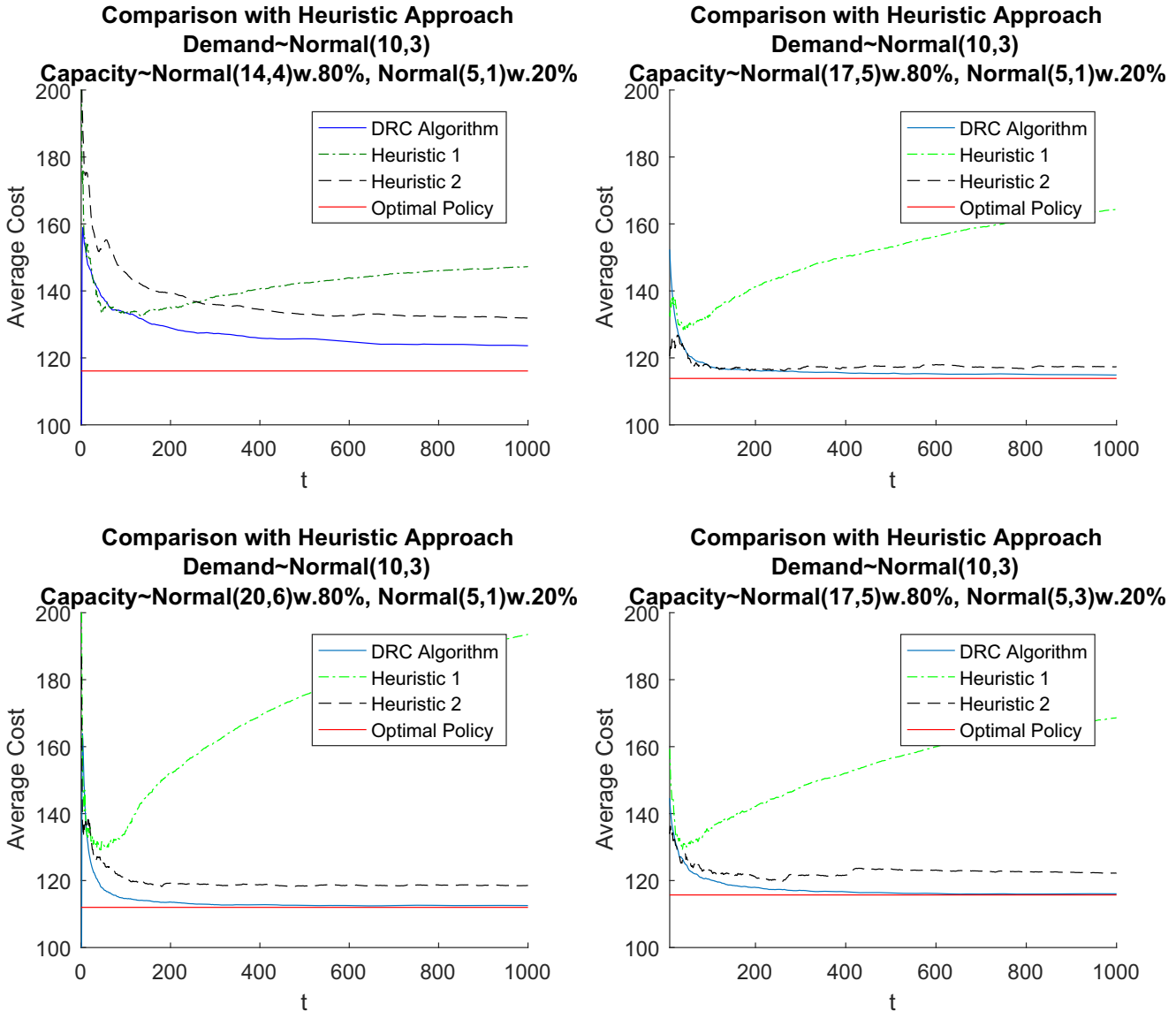
We conduct numerical experiments to demonstrate the efficacy of our proposed DRC algorithm. To the best of our knowledge, we are not aware of any existing learning algorithms that are applicable to random capacitated inventory systems. Thus, we have designed two simple heuristic learning algorithms (that are intuitively sound and practical), and use them as benchmarks to validate the performance of the DRC algorithm. Our results show that the performance of the DRC algorithms is superior to these two benchmarking heuristics both in terms of consistency and convergence rate. All the simulations were implemented on an Intel Xeon 3.50GHz PC.

6.1. Design of Experiments

We conduct our numerical experiments using a normal distribution for the random demand and a mixture of two normal distributions for the random capacity. More specifically, we set the demand to be $\mathbf{N}(10, 3^2)$. We test four different capacity distributions, namely, a mixture of 20% $\mathbf{N}(5, 1^2)$ and 80% $\mathbf{N}(14, 4^2)$, a mixture of 20% $\mathbf{N}(5, 1^2)$ and 80% $\mathbf{N}(17, 5^2)$, a mixture of 20% $\mathbf{N}(5, 1^2)$ and 80% $\mathbf{N}(20, 6^2)$, and also a mixture of 20% $\mathbf{N}(5, 3^2)$ and 80% $\mathbf{N}(17, 5^2)$. The distributions correspond to environments where the product capacity is subject to downtime. Clearly, in a production environment, capacity may be random even if no significant downtime occurs (e.g., due to variations in operator speed). However, machine downtime can significantly impact capacity. These examples correspond to situations where the production system experiences downtime that affects capacity with 20% probability. (We have experimented with other examples of downtime and obtained similar results.)

The production cost $c = 10$, and the salvaging value is set to be half of the production cost, that is, $\theta = 5$. The backlogging cost is linear in backorder quantity, with per-unit cost $b = 10$, and the holding cost is 2% per period of the production cost, that is, $h = 0.2$. We set the time horizon $T = 1000$, and compare the average cost of our DRC algorithm with that of the two benchmarking heuristic algorithms (described below) as well as the clairvoyant optimal cost over 1000 periods.

Figure 5 Computational Performance of the Data-Driven Random Capacity Algorithm (the average cost Case) [Color figure can be viewed at wileyonlinelibrary.com]



Clairvoyant Optimal Policy: The clairvoyant optimal policy is a stationary policy, given that the firm knows both the demand and capacity distributions at the beginning of the planning horizon. The average cost is calculated by averaging 1000 runs over 1000 periods.

Benchmarking Heuristic 1: We start with an arbitrary inventory level s_1 and start the first production cycle. For $t \geq 1$, we keep the target level $s_t = s_j$ the same during one production cycle $j \geq 1$. If the inventory level y_t reaches s_j, we claim that the j^{th} production cycle ends and then we collect all the past observed demand data to form an empirical demand distribution and all the past observed capacity data (except the capacity data obtained at the end of each production cycle) to form an empirical capacity distribution. We omit the capacity data obtained at the end of each production cycle because we might not produce at full capacity (when the previous target level is achieved). Then we treat the updated empirical demand and capacity distributions as true distributions, and derive the *long-run* optimal target level s_{j+1} for the subsequent cycle $j + 1$. Note that the long-run optimal target level (with well-defined input demand and capacity distributions) can be computed using the detailed computational procedure described in Ciarallo et al. (1994). The average cost is calculated by averaging 1000 runs over 1000 periods.

Benchmarking Heuristic 2: We start with an arbitrary inventory level s_1, and keep the target level $s_t = s_j$ the same during one production cycle $j \geq 1$. We still update the empirical demand distribution at the

end of each production cycle using all past observed demand data. However, in the first $N = 10$ periods, we always try to produce up to the maximum capacity $\bar{u}$, and we form the empirical capacity distribution using only these N full capacity sample points, and treat the empirical capacity distribution as the true capacity distribution for the rest of decision horizon. At the end of each production cycle, we still collect all the past observed demand data to form an empirical demand distribution, and similar to heuristic 1, derive the long-run optimal target level for the subsequent cycle together with the empirical capacity distribution. In other words, in the first N periods, we always produce up to the full capacity instead of the target level to get true information of the capacity, and after N periods, we carry out a regular modified base-stock policy. The average cost is calculated by averaging 1000 runs over 1000 periods. We have experimented with N values different than 10 and our results are similar to those we report below.

6.2. Numerical Results and Findings

The numerical results are presented in Figure 5. We observe that Heuristic 1 is inconsistent, that is, it fails to converge to the clairvoyant optimal cost. This is because even if we collect all the capacity data only when we produce at full capacity, the empirical distribution formed by these data is still biased (as the capacity data we observe is smaller than the true capacity). Heuristic 2 performs better than Heuristic 1, but still suffers from inconsistency.

Comparing to the benchmarking heuristic algorithms, the DRC algorithm converges to the clairvoyant optimal cost consistently and also at a much faster rate. We can also observe that when the capacity utilization (defined as the mean demand over the mean capacity) increases, the convergence rate slows down. This is because when the capacity utilization is high, it generally takes more periods for the system to reach the previous target level, resulting in longer production cycle length and slower updating frequency. Finally, we find that increasing the variability of distributions does not affect the performance of the DRC algorithm.

6.3. Extension to the Discounted Cost Case

We also conduct numerical experiments for the discounted cost case. More specifically, we choose the demand to be $\mathbf{N}(10, 3^2)$ and the production capacity to be a mixture of 20% $\mathbf{N}(5, 1^2)$ and 80% $\mathbf{N}(14, 4^2)$. The total cost can be written as $\sum_{t=1}^{T} \alpha^t \Omega(x_t, s_t)$ where $0 < \alpha < 1$ is the discount factor and $\Omega(x_t, s_t)$ is the single period cost. We compare our DRC algorithm with the optimal policy and two benchmarking heuristics under $\alpha = 0.995, 0.99, 0.97, 0.95$. The production, salvaging, backlogging, and holding costs are kept the same as the previous numerical experiment, that is, $c = 10$, $\theta = 5$, $b = 10$, $h = 0.2$. We compare the total cost up to $T = 1000$ periods. To adapt our DRC algorithm to the discounted cost case, we slightly modify our updating strategy in Step 1 as follows:

$$\mathcal{G}_k(\hat{s}_{\tau_j}) = \begin{cases} \alpha^{t-\tau_j} h, & \text{if } \hat{s}_{\tau_j} \wedge (\hat{x}_k + u_k) \geq d_k, \\ -\alpha^{t-\tau_j} b, & \text{otherwise.} \end{cases}$$

where $t - \tau_j$ is the time elapsed counting from the beginning of the current production cycle. The numerical results are presented in Figure 6. We observe that the DRC algorithm clearly outperforms the two benchmarking heuristics in terms of the total discounted cost.

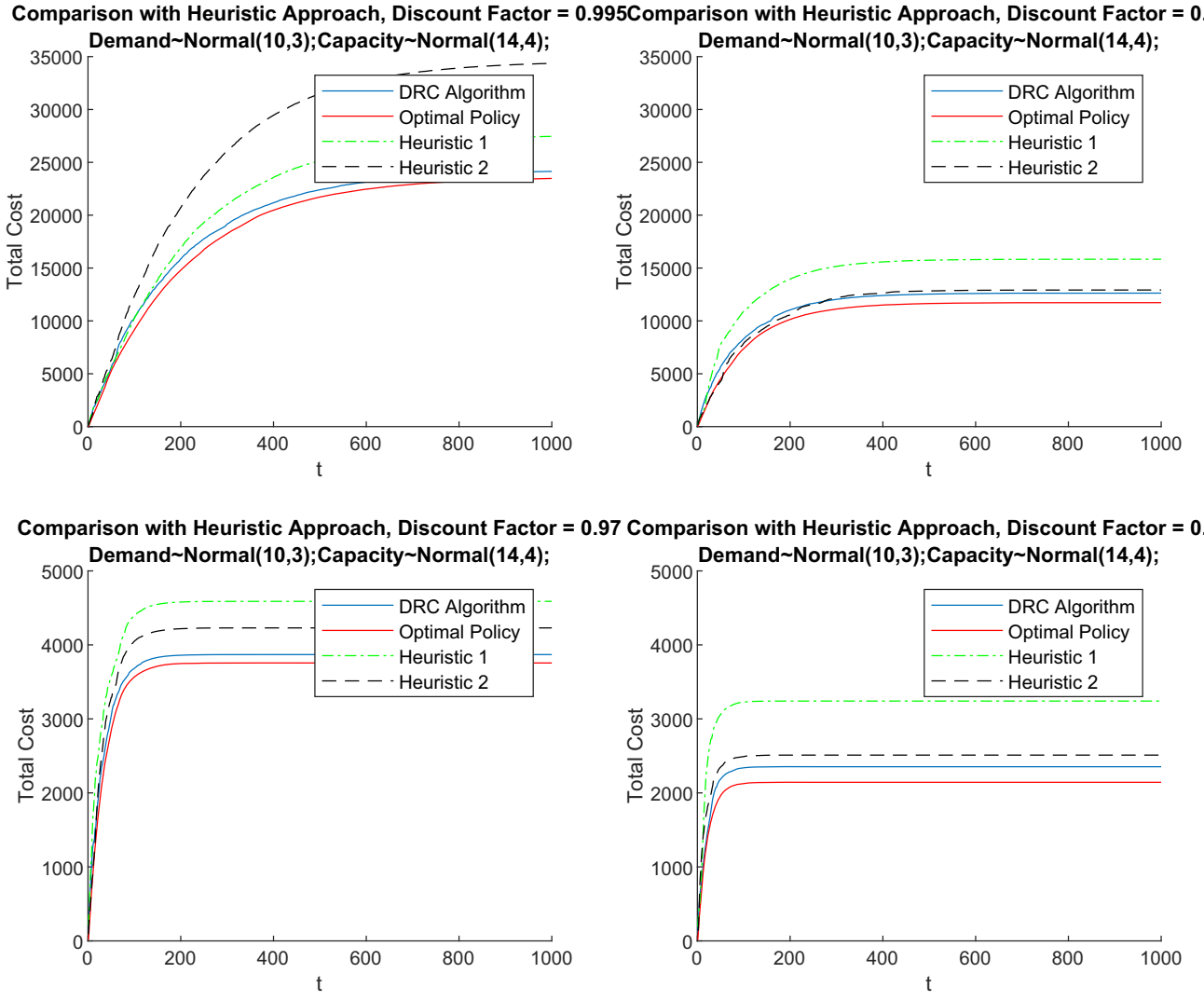
7. Concluding Remark

In this study, we have proposed a stochastic gradient descent type of algorithm for the stochastic inventory systems with random production capacity constraints, where the capacity is censored. Our algorithm utilizes the fact that the clairvoyant optimal policy is the extended myopic policy and updates the target inventory level in a cyclic manner. We have shown that the average T-period cost of our algorithm converges to the optimal cost at the rate of $O(1/\sqrt{T})$, which is the best achievable convergence rate. To the best of our knowledge, our study is the first paper to study learning algorithms for stochastic inventory systems under uncertain capacity constraints. We have also compared our algorithm with two straw heuristic algorithms that are easy to use, and we have shown that our proposed algorithm performs significantly better than the heuristics in both consistency and efficiency. Indeed, our numerical experiments have shown that with censored capacity information, the heuristics may not converge to the optimal policy.

We leave an important *open question* on how to design an efficient and effective learning algorithm for the capacitated inventory systems with lost-sales and censored demand. In this study, with backlogged demand, the length of the production cycle is independent of the target level, and therefore, the production cycles in our proposed algorithm and the optimal system are perfectly aligned. With lost-sales and censored demand, the length of the production cycle becomes dependent on the target level, and comparing any two feasible policies becomes much more challenging, which would require significantly new ideas and techniques.

Finally, we would also like to remark the connection between our online learning algorithm and deep reinforcement learning (DRL) algorithms. Needless to say, DRL is very popular nowadays and can be used to solve stochastic problems involving learning. We

Figure 6 Computational Performance of the Data-Driven Random Capacity Algorithm (the discounted cost case) [Color figure can be viewed at wile yonlinelibrary.com]



refer interested readers to the recent work by Gijsbrechts et al. (2019) that employed DRL in various inventory control settings. The major differences of DRL and our online learning algorithms are as follows: (1) DRL requires a vast amount of data at the beginning to build the deep neural network, and therefore is suitable for inventory system which has substantial amount of history data. On the other hand, our online learning algorithm assumes very limited information at the beginning, and learns to optimize from scratch. Second, DRL uses the stochastic gradient descent method to carry out backprorogation, but it is almost impossible to interpret how the decisions are made in each period. By contrast, our online learning algorithm is highly interpretable. Third, the efficiency and accuracy of DRL highly rely on the structure of deep neural network and the choice of hyper-parameters, which requires much crafting and fine-tuning. It is harder to obtain theoretical convergence results. Overall, we think that DRL is a very powerful method to solve complex problems where there is a substantial amount of data and the decision makers can accept the results from a black-box procedure.

Acknowledgment

The authors thank the Department Editor Professor Qi Annabelle Feng, the anonymous Senior Editor, and the anonymous referees for their very constructive and detailed comments, which have helped significantly improve both the content and the exposition of this study. The research is partially supported by NSF grant CMMI-1634505.

Appendix: Technical Proofs for Section 3

PROOF OF PROPOSITION 1. To prove the target interval policy, we write the optimal single-period cost function as follows.

$$\mathbb{E}[\Omega(x,s)] = \min\left\{\min_{s\geq x}\mathbb{E}[\Omega_+(x,s)], \min_{s<x}\mathbb{E}[\Omega_-(x,s)]\right\}, \tag{A1}$$

where

$$\begin{aligned}\mathbb{E}[\Omega_+(x,s)] = {} & c\cdot(1-F_U(s-x))(s-x)\\ & + c\cdot\int_0^{s-x} rf_U(r)dr + (1-F_U(s-x))\\ & \left[\int_s^{\infty} b(z-s)f_D(z)dz + \int_0^{s} h(s-z)f_D(z)dz\right]\\ & + \int_0^{s-x}\int_{x+r}^{\infty} b(z-x-r)f_D(z)dzf_U(r)dr\\ & + \int_0^{s-x}\int_0^{x+r} h(x+r-z)f_D(z)dzf_U(r)dr,\end{aligned} \tag{A2}$$

$$\begin{aligned}\mathbb{E}[\Omega_-(x,s)] = {} & \theta\cdot(s-x) + \left[\int_s^{\infty} b(z-s)f_D(z)dz\right.\\ & \left. + \int_0^{s} h(s-z)\,f_D(z)\,dz\right].\end{aligned} \tag{A3}$$

Notice that we produce up to s when $s \geq x$, and salvage down to s when $s < x$.

We shall explain that in (A2) we condition on the event $s \leq (x+U)$, which has a probability of $(1-F_U(s-x))$, we have $s \wedge (x+U) = s$ and apply the standard newsvendor integral $\mathbb{E}[s-D]^+ + \mathbb{E}[D-s]^+ = \int_0^s (s-z)dz + \int_s^{\infty}(z-s)dz$. Similarly conditioning on the event $s > (x+U)$, which has a probability of $F_U(s-x) = \int_0^{s-x} f_U(r)dr$, we have $s \wedge (x+U) = x+U$ and also apply the standard newsvendor integral. Allowing for salvaging, the target level s can always be achieved in (A3).

To show a *target interval policy* is optimal, we first show that (A2) and (A3) have global minimizers s_l^* and s_u^*, respectively. Then, we show that $0 \leq s_l^* \leq s_u^* < \infty$. Finally, we discuss different strategies based on different starting inventory levels to imply that a *target interval policy* is optimal.

By applying the Leibniz integral rule, the first partial derivative of (A2) with respect to s is

$$\begin{aligned}\frac{\partial}{\partial s}\mathbb{E}[\Omega_+(x,s)] = {} & (1-F_U(s-x))\\ & \left[c + \int_s^{\infty}\frac{\partial}{\partial s}b(z-s)\,f_D(z)\,dz + \int_0^{s}\frac{\partial}{\partial s}h(s-z)f_D(z)dz\right].\end{aligned}$$

It can be easily solved that the solution to the first-order optimality, denoted by s_l^*, is

$$s_l^* = F_D^{-1}\left(\frac{b-c}{h+b}\right)$$

and

$$c + \int_{s_l^*}^{\infty}\frac{\partial}{\partial s}b(z-s)f_D(z)dz + \int_0^{s_l^*}\frac{\partial}{\partial s}h(s-z)f_D(z)dz = 0. \tag{A4}$$

$(x,s)]/\partial s < 0$ for $s < s_l^*$, and $\partial\mathbb{E}[\Omega_+(x,q)]/\,\partial q > 0$ for $s > s_l^*$. Thus, we conclude that s_l^* is the global minimum of $\mathbb{E}[\Omega_+(x,s)]$.

Moreover, the second partial derivative of (A2) with respect to s is

$$\begin{aligned}& \frac{\partial^2}{\partial^2 s}\mathbb{E}[\Omega_+(x,s)]\\ & = cf_U(s-x) + (1-F_U(s-x))\left[\int_s^{\infty}\frac{\partial^2}{\partial^2 s}b(z-s)f_D(z)dz\right.\\ & \quad + \int_0^{s}\frac{\partial^2}{\partial^2 s}h(s-z)\,f_D(z)\,dz + f_D(s)(h+b)\int_0^{0}\Bigg]\\ & \quad -f_U(s-x)\left[\int_s^{\infty}\frac{\partial}{\partial s}b(z-s)\,f_D(z)\,dz\right.\\ & \quad \left. + \int_0^{s}\frac{\partial}{\partial s}h(s-z)\,f_D(z)\,dz\right]\\ & = (1-F_U(s-x))[(h+b)f_D(s)] - f_U(s-x)\\ & \quad [(h+b)F_D(s) - b + c].\end{aligned}$$

It is easy to see when $s \leq s_l^*$,

$$\begin{aligned}&(1-F_U(s-x))[(h+b)f_D(s)] > 0 \quad \text{and}\\ &f_U(s-x)[(h+b)F_D(s) - b + c] \leq 0.\end{aligned}$$

Therefore, when $s \leq s_l^*$, $\partial^2\mathbb{E}[\Omega_+(x,s)]/\partial s^2 \geq 0$, which suggests that $\mathbb{E}[\Omega_+(x,s)]$ is convex in $s \leq s_l^*$.

Similarly, the first partial derivative of (A3) with respect to s is

$$\frac{\partial}{\partial s}\mathbb{E}[\Omega_-(x,s)] = \theta + \int_s^{\infty} -bf_D(z)dz + \int_0^{s} hf_D(z)dz \tag{A5}$$

and it is straightforward to check

$$\frac{\partial^2}{\partial^2 s}\mathbb{E}[\Omega_-(x,s)] \geq 0,$$

be the solution to the first-order condition $\partial\mathbb{E}[\Omega_-(x,s)]/\partial s = 0$, and then the solution s_u^* is the global minimum of $\mathbb{E}[\Omega_-(x,s)]$.

Since $\theta \le c$, by comparing (A4) and (A5), we have $s_l^* \le s_u^*$. The optimal strategy is as follows.

1. When $s_l^* \le x \le s_u^*$, the firm decides to do nothing.
2. When $x < s_l^*$, the firm decides to produce up to s_l^* (as much as possible).
3. When $s_u^* < x$, the firm decides to salvage down to s_u^*.

The three cases discussed above can be readily illustrated in Figure 1. We sketch (A2) and (A3) as functions of $s = x + q$. The two curves are labeled "$q \ge 0$" and "$q < 0$," respectively. We note that (A2) and (A3) intersect at $q = 0$, as discussed earlier. The solid curve is the effective cost function $\Omega(s)$, which consists of the curve "$q \ge 0$" for $s \ge x$, and the curve "$q < 0$" for $s < x$.

Proof of Proposition 2. We first prove Proposition 2(a). Define $G_t^*(x_t)$ be the optimal cost from period t to period T with starting inventory x_t, then the optimality equation for the system can be written as follows.

$$G_t^*(x_t) \equiv \min\left\{\min_{s_t \ge x_t} G_{t+}(x_t, s_t), \min_{s_t < x_t} G_{t-}(x_t, s_t)\right\}, \quad \text{(A6)}$$

where

$$\begin{aligned} G_{t+}(x_t, s_t) &= \mathbb{E}[\Omega_+(x_t, s_t)] \\ &+ \int_0^\infty \int_0^{s_t - x_t} G_{t+1}^*(x_t + r - z) f_U(r) dr f_D(z) dz \\ &+ (1 - F_U(s_t - x_t)) \int_0^\infty G_{t+1}^*(s_t - z) f_D(z) dz, \end{aligned} \quad \text{(A7)}$$

$$G_{t-}(x_t, s_t) = \mathbb{E}[\Omega_-(x_t, s_t)] + \int_0^\infty G_{t+1}^*(s_t - z) f_D(z) dz, \quad \text{(A8)}$$

where $\mathbb{E}[\Omega_+(x_t, s_t)]$ and $\mathbb{E}[\Omega_-(x_t, s_t)]$ represent the cost functions of period t with the produce-up-to decision and the salvage-down-to decision, respectively, as in Proposition 1.

Our goal is to prove that a target interval policy is optimal for any period t, that is, there exist two threshold levels $s_{t,l}^*$ and $s_{t,u}^*$ such that the optimal target level s_t^* satisfies

$$s_t^* = \begin{cases} s_{t,l}^*, & x_t < s_{t,l}^*, \\ x_t, & s_{t,l}^* \le x_t \le s_{t,u}^*, \\ s_{t,u}^*, & x_t > s_{t,u}^*. \end{cases}$$

Lemma 8. *If $G_{t+1}^*(\cdot)$ is convex, then $G_t^*(\cdot)$ is also convex. Also, a target interval policy is optimal in period t.*

Proof. We first show that a target interval policy is optimal in period t. The cost function for period t consists of (A7) and (A8). When $s_t \ge x_t$, the cost function is (A7), and when $s_t < x_t$, the cost function is (A8). Since $G_{t+1}^*(\cdot)$ and $\mathbb{E}[\Omega_-(x_t, s_t)]$ are convex in s_t, then we have that (A8) is convex in s_t and we let $s_{t,u}^*$ be the global minimum for (A8). For (A7), the first-order condition is

$$\begin{aligned} \frac{\partial}{\partial s_t} G_{t+}(x_t, s_t) &= \frac{\partial}{\partial s_t} \mathbb{E}[\Omega_+(x_t, s_t)] + (1 - F_U(s_t \\ &\quad - x_t)) \int_0^\infty G_{t+1}^{*\prime}(s_t - z) f_D(z) dz \\ &= 0. \end{aligned} \quad \text{(A9)}$$

Let $s_{t,l}^*$ be the solution to (A9). Following the same arguments as in Proposition 1 and the convexity of $G_{t+1}^*(\cdot)$ and $\mathbb{E}[\Omega_+(x_t, s_t)]$ for $s_t \le s_{t,l}^*$, we conclude that $s_{t,l}^*$ is the global minimum for (A7). Also, since $\theta \le c$, we have that $s_{t,l}^* \le s_{t,u}^*$. Thus, a target interval policy is optimal by following the three cases discussed in the single-period problem in Proposition 1.

Next, we show that $G_t^*(x_t)$ is convex in x_t. Given $s_{t,l}^*$ and $s_{t,u}^*$, we can readily write $G_t^*(x_t)$ with respect to the starting inventory x_t as follows.

$$G_t^*(x_t) = \min\left\{\min_{s_t \ge x_t} G_{t+}(x_t, s_t), \min_{s_t < x_t} G_{t-}(x_t, s_t)\right\} =$$
$$\begin{cases} \mathbb{E}\left[\Omega_+(x_t, s_{t,l}^*)\right] + \int_0^\infty \int_0^{s_{t,l}^* - x_t} G_{t+1}^*(x_t + r - z) f_U(r) dr f_D(z) dz + (1 - F(s_{t,l}^* - x_t)) \int_0^\infty G_{t+1}^*(s_{t,l}^* - z_t) f_D(z) dz, & x_t < s_{t,l}^*, \\ \int_{x_t}^\infty b(z - x_t) f_D(z) dz + \int_0^{x_t} h(x_t - z) f_D(z) dz + \int_0^\infty G_{t+1}^*(x_t - z) f_D(z) dz, & s_{t,l}^* \le x_t \le s_{t,u}^*, \\ \mathbb{E}\left[\Omega_-(x_t, s_{t,u}^*)\right] + \int_0^\infty G_{t+1}^*(s_{t,u}^* - z) f_D(z) dz, & s_{t,u}^* < x_t, \end{cases} \quad \text{(A10)}$$

where $s_{t,l}^*$ and $s_{t,u}^*$ are the global minima defined earlier.

By the Leibniz integral rule, the second derivatives of (A10) with respect to x_t are

$$\frac{\partial^2}{\partial^2 x_t} G_t^*(x_t) = \begin{cases} \frac{\partial^2}{\partial^2 x_t}\mathbb{E}\left[\Omega_+(x_t, s_{t,l}^*)\right] + \int_0^\infty \int_0^{s_{t,l}^* - x_t} G_{t+1}^{*\prime\prime}(x_t + r - z) f_U(r) dr f_D(z) dz, & x_t < s_{t,l}^*, \\ (h+b) f_D(x_t) + \int_0^\infty G_{t+1}^{*\prime\prime}(x_t - z) f_D(z) dz, & s_{t,l}^* \le x_t \le s_{t,u}^*, \\ \frac{\partial^2}{\partial^2 x_t}\mathbb{E}\left[\Omega_-(x_t, s_{t,u}^*)\right] + \int_0^\infty G_{t+1}^{*\prime\prime}(s_{t,u}^* - z) f_D(z) dz, & s_{t,u}^* < x_t. \end{cases} \tag{A11}$$

Because $\mathbb{E}[\Omega_+(x_t, s_{t,l}^*)]$ and $\mathbb{E}[\Omega_-(x_t, s_{t,u}^*)]$ are convex (which has been derived in Proposition 1), and $G_{t+1}^{*\prime\prime}(\cdot)$ is positive (by the inductive assumption), we have that (A11) are all positive. This means that $G_t^*(x_t)$ is convex on these three intervals separately. It remains to show that $G_t^*(x_t)$ is convex on the entire domain by carefully checking the connecting points between these intervals. We have

$$\lim_{\delta \to 0^-} \frac{G_t^*(s_{t,l}^*) - G_t^*(s_{t,l}^* - \delta)}{\delta} = (h+b) F_D(s_{t,l}^*) - b + \int_0^\infty G_{t+1}^{*\prime}(s_{t,l}^* - z) f_D(z) dz,$$

$$\lim_{\delta \to 0^+} \frac{G_t^*(s_{t,l}^* + \delta) - G_t^*(s_{t,l}^*)}{\delta} = (h+b) F_D(s_{t,l}^*) - b + \int_0^\infty G_{t+1}^{*\prime}(s_{t,l}^* - z) f_D(z) dz,$$

$$\lim_{\delta \to 0^-} \frac{G_t^*(s_{t,u}^*) - G_t^*(s_{t,u}^* - \delta)}{\delta} = (h+b) F_D(s_{t,u}^*) - b + \int_0^\infty G_{t+1}^{*\prime}(s_{t,u}^* - z) f_D(z) dz,$$

$$\lim_{\delta \to 0^+} \frac{G_t^*(s_{t,u}^* + \delta) - G_t^*(s_{t,u}^*)}{\delta} = (h+b) F_D(s_{t,u}^*) - b + \int_0^\infty G_{t+1}^{*\prime}(s_{t,u}^* - z) f_D(z) dz.$$

Thus, we can see that the first derivatives at the connecting points are the same, and therefore $G_t^*(\cdot)$ is continuously differentiable and convex on the entire domain.

By definition, we know that $G_{T+1}^*(x_{T+1}) = -\theta(x_{T+1})$ is convex. Thus, by Lemma 8 and induction, we conclude that the target interval policy is optimal for any period $t = 1,\ldots,T$. This proves Proposition 2(a).

We then prove Proposition 2(b). The single-period cost and derivative are exactly the same for both the produce-up-to and salvage-down-to cases. The optimality equation for infinite horizon case can be written as

$$J(x) = \min\left\{\min_{s \ge x} G_+(x,s), \min_{s < x} G_-(x,s)\right\}.$$

where

$$G_+(x,s) = \mathbb{E}[\Omega_+(x,s)] + \alpha(1 - F(s-x)) \int_0^\infty J(s-z) f_D(z)\, dz, \tag{A12}$$

$$G_-(x,s) = \mathbb{E}[\Omega_-(x,s)] + \alpha \int_0^\infty J(s-z) f_D(z)\, dz, \tag{A13}$$

where $0 \le \alpha < 1$ is the discount factor. Our goal is to prove that a target interval policy is optimal, that is, there are two threshold levels s_l^* and s_u^* such that the optimal target level is s_l^* when $x < s_l^*$ and s_u^* when $x > s_u^*$ and x otherwise. Similar to Lemma 8, we can show that $J(x)$ is convex in the starting inventory x. The remainder argument is identical to that of Proposition 2(a). For the infinite horizon average cost problem, it suffices to verify the set of conditions in Schäl (1993), ensuring the limit of the discounted cost optimal policy is the average optimal policy as the discount factor $\alpha \to 1$ from the below. Verifying these conditions is a standard exercise in the literature, and thus we omit the details for brevity. This completes the proof.

References

Agrawal, S., R. Jia. 2019. Learning in Structured MDPS with Convex Cost Functions: Improved Regret Bounds for Inventory Management. Proceedings of the 2019 ACM Conference on Economics and Computation, EC '19, : ACM, New York, NY, USA, pp. 743–744.

Angelus, A., E. L. Porteus. 2002. Simultaneous capacity and production management of short-life-cycle, produce-to-stock goods under stochastic demand. *Management Sci.* **48**(3): 399–413.

Angelus, A., W. Zhu. 2017. Looking upstream: Optimal policies for a class of capacitated multi-stage inventory systems. *Prod. Oper. Manag.* **26**(11): 2071–2088.

Aviv, Y., A. Federgruen. 1997. Stochastic inventory models with limited production capacity and periodically varying parameters. *Probab. Engrg. Inform. Sci.* **11**: 107–135.

Ban, G. Y. 2020. Confidence intervals for data-driven inventory policies with demand censoring. Working paper, London Business School, London, UK.

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

Brownlee, J. 2014. Manufacturing Problems could Make the iPhone 6 Hard to Find at Launch. Available at #https://www.cultofmac. #com/285046/manufacturing-problems-make-iphone-6-hard-find-launch/ (accessed date October 29, 2018).

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

Chen, L., E. L. Plambeck. 2008. Dynamic inventory management with learning about the demand distribution and substitution probability. *Manuf. Serv. Oper. Manag.* **10**(2): 236–256.

Chen, B., C. Shi. 2020. Tailored base-surge policies in dual-sourcing inventory systems with demand learning. Working paper, University of Michigan, Ann Arbor, MI.

Chen, X., Gao, X., Z. Pang. 2018. Preservation of structural properties in optimization with decisions truncated by random variables and its applications. *Oper. Res.* **66**(2): 340–357.

Chen, B., Chao, X., H. S. Ahn (2019a). Coordinating pricing and inventory replenishment with nonparametric demand learning. *Oper. Res.* **67**(4): 1035–1052.

Chen, B., Chao, X., C. Shi. (2019b). Nonparametric algorithms for joint pricing and inventory control with lost-sales and censored demand. Working paper, University of Michigan, Ann Arbor, MI.

Chu, L. Y., Shanthikumar, J. G, Z. J. M. Shen. 2008. Solving operational statistics via a bayesian analysis. *Oper. Res. Lett.* **36**(1): 110–116.

Ciarallo, F. W., Akella, R., T. E. Morton. 1994. A periodic review, production planning model with uncertain capacity and uncertain demand — optimality of extended myopic policies. *Management Sci.* **40**(3): 320–332.

Duenyas I., Hopp, W. J., Y. Bassok. 1997. Production quotas as bounds on interplant JIT contracts. *Management Sci.* **43**(10): 1372–1386.

Eberly, J. C., J. A Van Mieghem. 1997. Multi-factor dynamic investment under uncertainty. *J. Econ. Theory* **75**(2): 345–387.

Federgruen, A., N. Yang. 2011. Procurement strategies with unreliable suppliers. *Oper. Res.* **59**(4): 1033–1039.

Federgruen, A., P. Zipkin (1986a). An inventory model with limited production capacity and uncertain demands I: The average-cost criterion. *Math. Oper. Res.* **11**(2): 193–207.

Federgruen, A., P. Zipkin (1986b). An inventory model with limited production capacity and uncertain demands II: The discounted cost criterion. *Math. Oper. Res.* **11**(2): 208–215.

Feng, Q. 2010. Integrating dynamic pricing and replenishment decisions under supply capacity uncertainty. *Management Sci.* **56**(12): 2154–2172.

Feng, Q., J. G. Shanthikumar. 2018. Supply and demand functions in inventory models. *Oper. Res.* **66**(1): 77–91.

Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., D. J. Zhang. 2019. Can deep reinforcement learning improve inventory management? performance on dual sourcing, lost sales and multi-echelon problems. Working paper, Northwestern University, Evanston, IL.

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 Sci.* **47**(8): 1101–1112.

Güllü, R. 1998. Base stock policies for production/inventory problems with uncertain capacity levels. *Eur. J. Oper. Res.* **105**(1): 43–51.

Hazan, E. 2016. Introduction to online convex optimization. *Found. Trends R Optimiz.* **2**(3-4): 157–325.

Henig, M., Y. Gerchak. 1990. The structure of periodic review policies in the presence of random yield. *Oper. Res.* **38**(4): 634–643.

Huh, W. T., M. Nagarajan. 2010. Linear inflation rules for the random yield problem: Analysis and computations. *Oper. Res.* **58** (1): 244–251.

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

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

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

Kapuscinski, R., S. Tayur. 1998. A capacitated production-inventory model with periodic demand. *Oper. Res.* **46**(6): 899–911.

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

Lariviere, M. A., E. L. Porteus. 1999. Stalking information: Bayesian inventory management with unobserved lost sales. *Management Sci.* **45**(3): 346–363.

Levi R., Roundy R. O., D. B. Shmoys. 2007. Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* **32**(4): 821–839.

Levi, R., Roundy, R. O., Shmoys, D. B., V. A. Truong. 2008. Approximation algorithms for capacitated stochastic inventory models. *Oper. Res.* **56**:1184–1199.

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

Liyanage, L. H., J. G. Shanthikumar. 2005. A practical inventory control policy using operational statistics. *Oper. Res. Lett.* **33**(4): 341–348.

Loulou, R. 1978. An explicit upper bound for the mean busy period in a GI/G/1 queue. *J. Appl. Probab.* **15**(2): 452–455.

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

Özer O.W. Wei. 2004. Inventory control with limited capacity and advance demand information. *Oper. Res.* **52**(6): 988–1000.

Powell, W., A., Ruszczyński, H., Topaloglu. 2004. Learning algorithms for separable approximations of discrete stochastic optimization problems. *Math. Oper. Res.* **29**(4): 814–836.

Randall, T., D. Halford. 2018. Tesla model 3 tracker. Available at https://www.bloomberg.com/graphics/2018-tesla-tracker/ (accessed date October 29, 2018).

Roundy, R. O., J. A. Muckstadt. 2000. Heuristic computation of periodic-review base stock inventory policies. *Management Sci.* **46**(1): 104–109.

Schäl, M. 1993. Average optimality in dynamic programming with general state space. *Math. Oper. Res.* **18**(1): 163–172.

Shalev-Shwartz, S. 2012. Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4**(2): 107–194.

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

Simchi-Levi, D., Chen, X., J. Bramel. 2014. *The Logic of Logistics: Theory, Algorithms, and Applications for Logistics and Supply Chain Management*, Springer, New York, NY.

Snyder, L. V., Z. J. M. Shen. 2011. *Fundamentals of Supply Chain Theory*, John Wiley & Sons, Hoboken, NJ.

Sohail, O. 2018. Production Problems Might Delay lcd iPhone 9 Model to Launch in November - Notch Said to be the Culprit. Available at https://wccftech.com/iphone-9-lcd-model-delayed-till-november/ (accessed October 29, 2018).

Sparks, D. 2018. Tesla model 3 production rate: 3000 units per week. Available at #https://finance.yahoo.com/news/ #tble-sla-model-3-production-rate-184600194.html (accessed October 29, 2018).

Tayur, S. 1992. Computing the optimal policy for capacitated inventory models. *Stoch. Models* **9**: 585–598.

Wang, Y., Y. Gerchak. 1996. Periodic review production models with variable capacity, random yield, and uncertain demand. *Management Sci.* **42**(1): 130–137.

Yuan, H., Luo, Q., C. Shi. 2019. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. Working paper, University of Michigan, Ann Arbor, MI.

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

Zhang, H., Chao, X., C. Shi. 2020. Closing the gap: A learning algorithm for the lost-sales inventory system with lead times. Working paper, University of Michigan, Ann Arbor, MI.

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