Data Driven Optimization:
Theory and Applications in Supply Chain Systems

by

Hao Yuan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2019

Doctoral Committee:

Assistant Professor Cong Shi, Chair
Assistant Professor Ruiwei Jiang
Associate Professor Siqian Shen
Assistant Professor Joline Uichanco

Hao Yuan
haoyuan@umich.edu
ORCID iD: 0000-0003-1738-8673

ACKNOWLEDGEMENTS

First, I express the deepest thanks to my advisor Professor Cong Shi. He went above and beyond in guiding and supporting me through my doctorate study. I appreciate the flexibility he allows and the great help he offers in both my study and life.

I would also like to thank my committee members Professor Siqian Shen, Professor Ruiwei Jiang and Professor Joline Uichanco for their helpful discussions and feedback.

My gratitude also goes to every professor and staff at the University of Michigan who taught or helped me during my Ph.D. journey, and to Professor Jonathan Smith, Professor Ling Long, Professor Mark Rudelson and Professor Sijue Wu who made me love mathematics.

Finally, I would like to acknowledge with much love and gratitude for my family and friends.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Supply chain optimization plays a critical role in many business enterprises. In a data driven environment, rather than pre-specifying the underlying demand distribution and then optimizing the system's objective, it is much more robust to have a nonparametric approach directly leveraging the past observed data. In the supply chain context, we propose and design *online learning algorithms* that make adaptive decisions based on historical sales (a.k.a. censored demand). We measure the performance of an online learning algorithm by *cumulative regret* or simply *regret*, which is defined as the cost difference between the proposed algorithm and the clairvoyant optimal one.

In the presence of inventory constraints and censored demand information, the data driven supply chain optimization falls into the broad domain of reinforcement learning. To design learning algorithms with theoretical performance guarantees, in Chapter 1, we first discuss a general framework that uses the so-called *cycling-trick* to transform the reinforcement learning problem into a variant of the multi-armed bandit problem (MAB), which we refer to as cyclic online learning, and design an upper confidence bound (UCB) type algorithm that achieves the optimal regret rate.

In the online learning literature, the most popular algorithms are *stochastic gradient descent* (SGD) based, which leverages the convexity properties in the objective functions. In Chapter 2, we use the newsvendor problem with fixed cost as an example to demonstrate that some new online learning methods can be applied when the objective is not convex and the naive SGD algorithm fails. In one method, we consider the *bandits on convex function* problem, and design an algorithm that combines both the first order method (i.e., stochastic gradient descent) and the zeroth order method (i.e., multi-armed bandits control). In another method, we design a pure zeroth order method, termed "Shrinking Active Set (SAS) algorithm", that exploits the "one-side information" revealed by past sales. We show that both algorithms achieve provably optimal regret rate. In Chapter 3, we further extend the SAS algorithm to a Markovian environment. More specifically, we consider the lost-sales inventory control problem with positive lead time, and show that if the system dynamics under a fixed policy has the uniform ergodic property, then the SAS algorithm can be applied to achieve the optimal regret rate.

In Chapter 4, we consider the multi-product inventory control problem, where the regret rate achieved by a stochastic gradient type algorithm is optimal with respect to the time horizon T but sub-optimal with respect to the number of products J. With a new space transformation, we design a *mirror decent* type algorithm that improves the known regret rate by a factor of $\sqrt{J/\log(J)}$.

In Chapter 5, we consider the periodic-review inventory control problem with fixed cost. To achieve the optimal regret rate, we combine many techniques developed above: cyclic online learning, stochastic gradient descent, bandits on convex function and shrinking active set algorithm.

In the supply chain context, to design efficient learning algorithms, we typically face two major challenges. First, we need to identify a suitable recurrent state that decouples system dynamics into cycles with good properties: (1) smoothness and rich feedback information necessary to apply the zeroth order optimization method effectively; (2) convexity and gradient information essential for the first order methods. Second, we require the learning algorithms to be adaptive to the *physical constraints*, e.g., positive inventory carry-over, warehouse capacity constraint, ordering/production capacity constraint, and these constraints limit the policy search space in a dynamic fashion. To design efficient and provably-good data driven supply chain algorithms, we zoom into the detailed structure of each system, and carefully trade off between exploration and exploitation.

CHAPTER 1

Cyclic Online Learning

1.1 Introduction

A supply chain optimization problem can often be seen as a Markov decision process with inventory level as states, ordering quantities as decisions, and inventory cost as cost. The uncertainty in the demand results in random state transitions. We study data driven models where the demand distribution is unknown, and design reinforcement learning algorithms to learn demand and optimize ordering decisions. We measure the performance of learning algorithms by *regret*: the cost difference between its derived policy and optimal policy.

In recent works (Chen et al. (2018c); Zhang et al. (2018, 2019)), a popular approach is to use a so-called *cycling trick* that groups periods into cycles so that we evaluate a policy according to its cycle performance. On a high level, the cycling trick transforms the data driven supply chain optimization problem from a reinforcement learning problem into an online learning problem, where, to achieve a theoretical regret bound, people usually apply two families of online learning methods: multi-armed bandits control and online convex optimization.

More concretely, consider a supply chain policy π mapping inventory level to ordering decision. The inventory levels over time forms a Markov chain. Under mild regularity conditions, this Markov chain contains a recurrent state r and a cycle as a group of periods from one hitting period of r to the next. We can prove that the long run average cost is equal to the cycle average cost:

$$\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}C_t^{\pi}\right]=\frac{\mathbb{E}\left[G^{\pi}\right]}{\mathbb{E}\left[L^{\pi}\right]} \tag{1.1}$$

where LHS is the long run average cost with C_t^{π} denoting the inventory cost at period t for policy π, and RHS is the cycle average cost with G^{π} and L^{π} denoting random cycle cost and cycle length for a policy π, respectively.

When demand distribution is unknown, a learning algorithm provides a sequence of policy $\{\pi_t\}_{t=1}^{\infty}$ where π_t depends on observed information prior to period t. The performance is measured by *regret* formulated as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi_t} - v^*\right] \tag{1.2}$$

where

$$v^* = \min_{\pi}\left(\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi}\right]\right)$$

is the minimum long run average cost. This form of $\mathcal{R}_T$ is hard to analyze because $C_t^{\pi_t}$ is not a good evaluation of policy π_t. It has so-called *delayed cost issue*: A policy π_t that gives a good immediate cost $C_t^{\pi_t}$ may lead to a bad next-period state x_{t+1} that compromises future cost. On the other hand, if π_t is updated cycle by cycle, the standard regret can be transformed into *a cycle regret* as follows:

$$\begin{aligned}\mathcal{R}_N &= \mathbb{E}\left[\sum_{t=1}^{T}\left(C_t^{\pi_t} - v^*\right)\right]\\ &= \mathbb{E}\left[\sum_{n=1}^{N}\left(G_n^{\pi_n} - L_n^{\pi_n} v^*\right)\right]\end{aligned}$$

where T is the number of periods in N cycles. Note that the beginning state of each cycle is the same state r. This helps avoid the *delayed cost issue*. The problem then becomes an online learning problem: We update policy π_n cycle by cycle with previous cycles' information. Unlike traditional period-by-period online learning where we only consider the cost, here we account for both cycle cost $G_{n'}^{\pi_{n'}}$ and cycle length $L_{n'}^{\pi_{n'}}$ for all $n' < n$. We call this cycle-by-cycle online learning approach the *cyclic online learning.*

It is worth mentioning that, in previous cyclic online learning problems motivated in supply chain optimization (see Shi et al. (2016); Zhang et al. (2018, 2019)), their policies are independent to the random cycle length, i.e., $\mathbb{E}[L^{\pi}] = \mathbb{E}[L]$ for all policy π. In this case, by defining $\tilde{v} := \mathbb{E}[L]\, v^*$, the cycle regret can be transformed into

$$\mathcal{R}_N = \mathbb{E}\left[\sum_{n=1}^{N}\left(G_n^{\pi_n} - \tilde{v}^*\right)\right]. \tag{1.3}$$

where we only need to find policy that minimizes cycle cost G^{π}, making it the same form of classic online learning where cycle length plays no role. In supply chain optimization problems, cycle length L^{π} indeed depends on policy π. Well-known examples are the s-S policy with recurrent state $r = S$ (Scarf (1960)) and the single index policy with recurrent state with r as regular order-up-to parameter (Scheller-Wolf et al. (2007)). In this work, we present a formulation of the cyclic online learning problem and propose a learning algorithm that achieves optimal regret rate.

1.2 Cyclic Online Learning and CLCB Algorithm

1.2.1 Model

We model the *cyclic online learning* as a variation of the discrete online learning problems (aka multi-armed bandit problems): In each cycle $n = 1,2\ldots$, there are random cycle cost function $G_n : [J] \to \mathbb{R}^+$ and random cycle length function $L_n : [J] \to \mathbb{Z}^+$, where J is the total number of actions. Note that L_n takes a positive integer value, since it refers to the number of periods in a cycle.

We assume G_n and L_n are possibly dependent, but, across each cycle, the sequence $\{(G_n, L_n)\}_{n=1}^{\infty}$ are i.i.d. random variables. We use (G, L) to denote the time generic random functions (G_n, L_n) for $n = 1,2\ldots$ i.e., $(G,L) = (G_n, L_n)$ in distribution. The distribution of (G,L) are unknown to us. At each cycle $n = 1,2,\ldots$, we select a policy $j_n \in [J]$, and observe the cycle cost $G_n^{j_n}$ and cycle length $L_n^{j_n}$. To lighten notation, we use $v^j := \frac{\mathbb{E}[G^j]}{\mathbb{E}[L^j]}$ to denote the cycle average cost for policy j. The optimal policy j^* minimizes the cycle average cost. i.e., $j^* = \underset{j\in[J]}{\arg\min} v^j$. We denote $v^* := v^{j^*}$. The goal is to minimize the *cyclic regret*:

$$\mathcal{R}_N := \mathbb{E}\left[\sum_{n=1}^{N}\left(G_n^{j_n} - L_n^{j_n} v^*\right)\right], \tag{1.4}$$

Within N cycles, $\mathbb{E}\left[\sum_{n=1}^{N} G_n^{j_n}\right]$ is the expected total cost given the sequence of policies $j_1, j_2, \ldots, j_N$, and $\mathbb{E}\left[\sum_{n=1}^{N} L_n^{j_n} v^*\right]$ is the expected total cost assuming we suffer the optimal cost v^* in every period. By the definition of v^*, it is evident that $\mathbb{E}\left[G_n^{j_n} - L_n^{j_n} v^*\right] \geq 0$ for any j^n, so $\mathcal{R}_N$ is a growing function N. We use big-O notation to measure its growing rate, and want to design learning algorithm with optimal rate.

Traditional works assume cycle length is independent of policies and the regret can be written as (1.3). Moreover, for distribution regularity condition, the vast majority of authors assume that, for each policy j, the unknown cost distributions are sub-gaussian, that is, the moment generating function of each G^j is such that then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda(G^j - \mathbb{E}G^j)} \leq e^{\sigma^2\lambda^2/2},$$

where $\sigma > 0$, the sub-gaussian parameter that is usually assumed to be known. In particular, if rewards take values in [0, 1], then by Hoeffding's lemma, one may take $\sigma = 1/4$. One can show that the UCB strategy (Auer et al. (2002)) has the following distribution independent regret rate:

$$\mathcal{R}_N = O\left(\sqrt{\log(N)NJ}\right).$$

We refer the reader to (Bubeck and Cesa-Bianchi (2012)) for a survey of the extensive literature in this area.

In cyclic online learning, cycle length L^j depends on policy j, and we make the following assumptions.

Assumption 1. *We assume:*

1. *For each policy j, G^j and L^j satisfy $G^j/L^j < \gamma$, for some known positive constant γ.*
2. *Marginally, both distribution of G^j and distribution of L^j are sub-exponential with parameters (ν, b).*

In supply chain optimization problems, a cycle extends the concept of period with one cycle consists of random number of periods. The first assumption holds if the cost in every period is bounded by γ, which, given a bounded inventory capacity constrain, is almost always the case. The second sub-exponential assumption is weaker than the traditional sub-gaussian assumption, which makes our estimation harder. But, this assumption is necessary: In most cases, the cycle length counts the number of periods until demand in the cycle accumulates to be above some positive threshold, which is usually sub-exponential but not sub-gaussian. For a simple example, consider D is the Bernoulli distribution ($\mathbb{P}[D=1] = \mathbb{P}[D=0] = 1/2$), then the hitting time of a positive level is geometric distribution (which is sub-exponential but not sub-gaussian).

If cycle length is independent of policy, the best policy has minimum expected cycle cost: $j^* = \arg\min_j \mathbb{E}\left[G^j\right]$. The key step for a learning algorithm is to estimate mean cycle cost $\mathbb{E}G^j$ for $j \in [J]$. In traditional UCB algorithm (Bubeck and Cesa-Bianchi (2012)), with n samples of G^j, $\left\{G^j_s\right\}_{s=1}^n$, the empirical average $\frac{1}{n}\sum_{s=1}^n G^j_s$ is an unbiased estimation of $\mathbb{E}G^j$, so we only need to calculate estimation error due to *variance*. On the other hand, in cyclic online learning, the cycle length L^j depends on j. The best policy minimizes cycle average cost: $j^* = \arg\min_j \frac{\mathbb{E}\left[G^j\right]}{\mathbb{E}\left[L^j\right]}$. A good learning algorithm has to estimate $\frac{\mathbb{E}\left[G^j\right]}{\mathbb{E}\left[L^j\right]}$. Compared with $\mathbb{E}G^j$, our objective $\frac{\mathbb{E}\left[G^j\right]}{\mathbb{E}\left[L^j\right]}$ is a nonlinear function of expectations. With n data $\left\{G^j_s\right\}_{s=1}^n$ drawn i.i.d. according to distribution of G^j and $\left\{L^j_s\right\}_{s=1}^n$ drawn i.i.d. according to distribution of L^j. The nonlinear version of empirical estimator

$$\hat{V}^j_n := \frac{\sum_{s=1}^n G^j_s}{\sum_{s=1}^n L^j_s} \tag{1.5}$$

is indeed biased: $\mathbb{E}\left[\hat{V}_n^j\right] \neq v^j$. We call $\hat{V}_n^j$ the *cycle empirical estimator.* In this paper, we design data driven policy that computes confidence interval for biased estimator $\hat{V}_n^j$, which handles error due to both bias and variance. We call it Cycle Lower Confidence Bound (CLCB) algorithm, and prove that it has regret rate

$$\mathcal{R}_N = O\left(\log N\sqrt{JN}\right).$$

1.2.2 Cycle Lower Confidence Bound Policy

The idea behind lower confidence bound (lcb) strategies (see Lai and Robbins (1985), Agrawal (1995) and Auer et al. (2002)) is that one should choose an arm for which the sum of its estimated mean and a confidence interval is lowest. When the cost distributions all satisfy the sub-Gaussian condition for a parameter σ, then such a confidence interval is easy to obtain. Suppose that at a certain time instance arm j has been sampled n times and the observed costs are $G_1^j,...,G_n^j$. Then the $G_1^j,...,G_n^j$ are i.i.d. random variables with mean $\mathbb{E}G^j$ and by a simple Chernoff bound, for any $\delta \in (0,1)$, the empirical mean $\frac{1}{n}\sum_{i=1}^n G_i^j$ satisfies, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{s=1}^n G_s^j \geq \mathbb{E}G^j - \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}}. \tag{1.6}$$

This property of the empirical mean turns out to be crucial in order to achieve a regret of optimal order.

The key of handling cyclic online learning is to replace the empirical mean with the so-called cycle empirical *estimator* $\hat{V}_n^j$ defined in (1.5). We need a similar performance guarantee like the one shown above for the empirical mean as in (1.6). More precisely, we want our cycle empirical estimator have the following property.

Proposition 1.1. *For any $j \in [J]$ and $n = 1,2,\ldots,$ we have, with probability at least* $1-\delta$,

$$\hat{V}_n^j \geq v^j - \theta\frac{\log(4/\delta)}{\sqrt{n}}$$

where

$$\theta := (2\gamma+2)\max(v,b). \tag{1.7}$$

Proof. From the definition 1.5, $\hat{V}_n^j$ is computed based on n samples of G^j and n samples of L^j. Let $\hat{G}_n^j$ and $\hat{L}_n^j$ denote empirical mean of these samples:

$$\hat{G}_n^j = \frac{1}{n}\sum_{s=1}^n G_s^j \text{ and } \hat{L}_n^j = \frac{1}{n}\sum_{s=1}^n L_s^j.$$

Consider

$$\begin{aligned}\left|\hat{V}_n^j - v^j\right| &= \left|\frac{\hat{G}_n^j}{\hat{L}_n^j} - \frac{\mathbb{E}G^j}{\mathbb{E}L^j}\right| = \left|\frac{\hat{G}_n^j}{\hat{L}_n^j} - \frac{\hat{G}_n^j}{\mathbb{E}L^j} + \frac{\hat{G}_n^j}{\mathbb{E}L^j} - \frac{\mathbb{E}G^j}{\mathbb{E}L^j}\right| \\ &= \frac{\hat{G}_n^j}{\hat{L}_n^j \mathbb{E}L^j}\left|\mathbb{E}L^j - \hat{L}_n^j\right| + \frac{1}{\mathbb{E}L^j}\left|\hat{G}_n^j - \mathbb{E}G\right| \le \gamma \left|\mathbb{E}L^j - \hat{L}_n^j\right| + \left|\hat{G}_n^j - \mathbb{E}G^j\right| \end{aligned} \tag{1.8}$$

where the inequality is due to $G^j/L^j \le \gamma$ and $\mathbb{E}L^j \ge 1$. Because G^j and L^j are (v,b) sub-exponential, with probability at least $1-\delta/2$, we have

$$\left|\hat{G}_n^j - \mathbb{E}G^j\right| < \max\left(\sqrt{\frac{2v^2 \log(4/\delta)}{n}}, \frac{2b\log(4/\delta)}{n}\right),$$

and, with probability at least $1-\delta/2$, we have

$$\left|\hat{L}_n^j - \mathbb{E}L^j\right| < \max\left(\sqrt{\frac{2v^2 \log(4/\delta)}{n}}, \frac{2b\log(4/\delta)}{n}\right).$$

Plugging these two sub-exponential concentration inequalities into (1.8), we get, with probability at least $1-\delta$,

$$\begin{aligned}\left|\hat{V}_n^j - v^j\right| &\le \gamma \max\left(\sqrt{\frac{2v^2 \log(4/\delta)}{n}}, \frac{2b\log(4/\delta)}{n}\right) + \max\left(\sqrt{\frac{2v^2 \log(4/\delta)}{n}}, \frac{2b\log(4/\delta)}{n}\right) \\ &\le (2\gamma+2)\max(v,b)\frac{\log(4/\delta)}{\sqrt{n}}\end{aligned}$$

This gives us the desired result. □

Now, we describe our cycle lower confidence bound algorithm (CLCB) as in Algorithm 1.1. We denote by T_n^j the (random) number of times policy j is selected up to cycle n.

The following proposition gives a regret bound for the CLCB algorithm.

Proposition 1.2. *The regret of the CLCB policy satisfies*

$$\mathcal{R}_N = O\left(\log N \sqrt{NJ}\right). \tag{1.9}$$

Proof. Let Δ^j denote the sub-optimality gap of playing policy j. i.e.,

$$\Delta^j := v^j - v^*.$$

Regret is the sum of loss due to playing suboptimal policies. We first bound the expected number

Algorithm 1.1 Cycle lower confidence bound algorithm(CLCB)

For each policy j, define cycle empirical estimator $\hat{V}_i^j$ as in (1.1) based on the first i observed cycle cost/length pairs $(G_1, L_1), \dots, (G_i, L_i)$ for policy j. Define the lower bound

$$B_{n,i}^j = \begin{cases} \hat{V}_n^j - \frac{\theta \log(4n^4)}{\sqrt{i}} & \text{for } n \geq 1 \\ -\infty & \text{for } n = 0, \end{cases}$$

where θ is defined in (1.7).

At cycle n, draw the arm minimizing $B_{n,T_{n-1}^j}^j$.

of pulls for a suboptimal arms. More precisely, in the first two steps of the proof we prove that, for any j such that $\Delta^j > 0$,

$$\mathbb{E}\left[T_N^j\right] \leq \frac{4\theta^2}{(\Delta^j)^2}\left(\log\left(4N^4\right)\right)^2 + 5 \tag{1.10}$$

First step.

Let cycle n, j_n denotes the policy we played at cycle n. We show that if $j_n = j$, then one of the following three inequalities is true:

$$B_{n,T_{n-1}^{j^*}}^{j^*} \geq v^*, \tag{1.11}$$

$$\hat{V}_{T_{n-1}^j}^j < v^j - \frac{\theta \log\left(4n^4\right)}{\sqrt{T_{n-1}^j}}, \tag{1.12}$$

$$T_{n-1}^j < \frac{4\theta^2}{(\Delta^j)^2}\left(\log\left(4N^4\right)\right)^2. \tag{1.13}$$

Indeed, assume that all three inequalities are false. Then we have

$$B_{n,T_{n-1}^{j^*}}^{j^*} < v^* = v^j - \Delta^j \leq v^j - \frac{2\theta \log\left(4N^4\right)}{\sqrt{T_{n-1}^j}} \leq \hat{V}_{T_{n-1}^j}^j - \frac{\theta \log\left(4n^4\right)}{\sqrt{T_{n-1}^j}} = B_{n,T_{n-1}^j}^j$$

which implies, in particular, that $j_n \neq j$.

Second step.

Here we first bound the probability that (1.11) or (1.12) hold. By preposition 1.1 as well as an union bound over the value of $T^{j^*}_{n-1}$ and T^{j}_{n-1} we obtain

$$\mathbb{P}\left[(1.11) \text{ or } (1.12) \text{ is true}\right] \leq 2\sum_{s=1}^{n}\frac{1}{s^4} \leq \frac{2}{n^3}.$$

Now, let

$$u := \left\lceil \frac{4\theta^2}{(\Delta^j)^2}\left(\log\left(4N^4\right)\right)^2 \right\rceil.$$

Using the first step, we obtain

$$\begin{aligned}\mathbb{E}\left[T^j_N\right] = \mathbb{E}\sum_{n=1}^{N}\mathbf{1}_{\{j_n=j\}} &\leq u + \mathbb{E}\sum_{n=u+1}^{N}\mathbf{1}_{\{j_n=j \text{ and } (1.13) \text{ is false}\}} \\ &\leq u + \mathbb{E}\sum_{n=u+1}^{N}\mathbf{1}_{\{j_n=j \text{ and } (1.11 \text{ or } 1.12) \text{ is true}\}} \leq u + \sum_{n=u+1}^{N}\frac{2}{n^3} \leq u+4\end{aligned}$$

This concludes the proof of (1.10).

Third Step.

Definition of cyclic regret (1.4) implies $\mathcal{R}_N = \sum_{j=1}^{J}\Delta^j\mathbb{E}\left[T^j_N\right]$. By (1.10), we immediately obtain

$$\mathcal{R}_N \leq \sum_{j:\Delta^j>0}\left(\frac{4\theta^2}{\Delta^j}\left(\log\left(4N^4\right)\right)^2 + 5\Delta^j\right).$$

Then to obtain (1.9), we use Hölder's inequality. For $N \geq n_0$ with some n_0 such that $\min_j \frac{4\theta^2}{(\Delta^j)^2}\left(\log\left(4N^4\right)\right)^2 \geq 5$,

$$\begin{aligned}\mathcal{R}_N = \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T^j_N}\sqrt{\mathbb{E}T^j_N} &\leq \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T^j_N}\sqrt{\frac{4\theta^2}{(\Delta^j)^2}(\log(4N^4))^2+5} \\ &\leq \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T^j_N}\sqrt{\frac{8\theta^2}{(\Delta^j)^2}(\log(4N^4))^2} \leq 2\sqrt{2}\theta\log\left(4N^4\right)\sum_{j:\Delta^j>0}\sqrt{\mathbb{E}T^j_N} \\ &\leq 2\sqrt{2}\theta\log\left(4N^4\right)\sqrt{\sum_{j:\Delta^j>0}\mathbb{E}T^j_N}\sqrt{\sum_{j:\Delta^j>0}1} = 2\sqrt{2}\theta\log\left(4N^4\right)\sqrt{NJ} = O\left(\log N\sqrt{NJ}\right).\end{aligned}$$

where the second inequality is due to $N \geq n_0$ and the last inequality is by applying Hölder's inequality. □

CHAPTER 2

New Online Learning Methods for Newsvendor Problem with Fixed Cost

In the majority of data driven supply chain optimization problems (Chen et al. (2018c); Shi et al. (2016); Zhang et al. (2018, 2019)), the solo unknown information is the demand distribution. If demand is uncensored which means, however making the ordering decisions, we always observe the realized demands D_t for $t = 1, 2 \ldots$, the most popular learning algorithm is the so-called sampled average approximation algorithm (SAA) (Levi et al. (2007)): At period t, we approximate the true demand distribution with empirical distribution $\hat{F}_t$ and apply optimal policy with respect $\hat{F}_t$, where , with respect to some suitable norm, this approximation has $1/\sqrt{t}$ convergence rate,

$$\left\|\hat{F}_t - F\right\| = O\left(1/\sqrt{t}\right).$$

Under mild regularity condition, SAA leads to regret at t of order $O\left(1/\sqrt{t}\right)$, and, therefore, its total regret is $O\left(\sqrt{T}\right)$.

SAA gives a universal solution for the data driven supply chain optimization with uncensored demand information. For the problem with censored demands, the algorithm design is more challenging. The dominating approaches are based on stochastic gradient descend (SGD) (see Chen et al. (2018c); Huh and Rusmevichientong (2009); Shi et al. (2016)), the standard online convex optimization theory (see Hazan et al. (2016)) implies the regret is also $O\left(\sqrt{T}\right)$. But, for many problems, due to the lack of either convexity of objective or the stochastic gradient information, the SGD approaches are hard to apply. In this paper, we seek for alternative approaches for handling the problems with censored demands, which maintain almost the same regret rate: $O\left(\sqrt{\log(T)T}\right)$.

As a concrete example, we consider the repeated newsvendor problem. With notation c: unit ordering cost, p : selling price, q_t : ordering quantity at period t, and D_t : random demands at period t for $t = 1, 2, \ldots$, the Newsvendor's cost is given by ordering cost minus sales revenue:

$$C_t := cq_t - p\min(q_t, D_t).$$

We assume $\{D_t\}_{t=1}^{\infty}$ are i.i.d. random variables, and use D to denote a time generic demand. We note that

$$\nabla C_t = c - p\mathbf{1}_{\{q_t<D_t\}},$$

so

$$\nabla\mathbb{E}\left[C_t\right] = c - p\mathbb{P}\left[q_t < D_t\right]$$

The optimal ordering quantity is $q^* = F^{-1}\left(1-\frac{c}{p}\right)$, where F denotes the cdf of demand D. But, in practice, F is unknown. With a learning algorithm, at period t, we choose ordering quantity q_t according to previous information to minimize *regret*

$$\mathcal{R}_T = \mathbb{E}\left[C_t(q_t) - C_t(q^*)\right].$$

With standard assumptions, we put an ordering capacity constrain $q_t \leq \beta$ for some positive constant β, and let $c < p$ (otherwise, ordering noting is obviously optimal).

The SAA approach is to order $q_t := \hat{F}_t\left(1-\frac{c}{p}\right)$, where $\hat{F}_t$ is the empirical distribution formed by previous demands $d_1\ldots,d_{t-1}$, which achieves $O\left(\sqrt{T}\right)$ regret. But, in practice, we only observe the sales data $\min\left(q_t,D_t\right)$, which leads on demand censoring and make the SAA approach invalid. To overcome this, the SGD approach updates $q_{t+1} = \mathbf{Proj}_{[0,\beta]}\left(q_t - \eta_t\tilde{\nabla}_t\right)$ where

$$\tilde{\nabla}_t := \nabla C_t = c - p\mathbf{1}_{\{q_t<D_t\}}$$

is an stochastic gradient for the expected cost $\mathbb{E}\left[C_t\right]$, which can be computed using sales information only. Since $\mathbb{E}\left[C_t(q_t)\right]$ is a convex function of q_t, standard online convex optimization theory implies $O\left(\sqrt{T}\right)$ regret with suitable choice of step size η_t (see Hazan et al. (2016)).

Now, We consider a modification of the Newsvendor problem. If $q_t \geq Q$ for some positive Q, the newsvendor has to pay a fixed cost K for renting a car for picking up his orders, which leads to extra fixed cost $K\mathbf{1}_{\{q_t\geq Q\}}$ on our cost. i.e., the new cost

$$\tilde{C}_t := K\mathbf{1}_{\{q_t\geq Q\}} + cq_t - p\min\left(q_t,D_t\right). \tag{2.1}$$

We call this problem *the News-vendor problem with Fixed Cost*. Note that

$$\tilde{\nabla}_t := c - q\mathbf{1}_{\{q_t<D_t\}} \tag{2.2}$$

is still a stochastic gradient of $\mathbb{E}\left[\tilde{C}_t\right]$, but the standard SGD algorithm cannot achieve $O\left(\sqrt{T}\right)$, since the objective $\mathbb{E}\left[C_t\right]$ is no longer convex. For general non-convex function, the first order methods, updating each step with gradient information, has no performance guarantee. Instead, we seek solutions from zeroth order methods: estimating objective directly from noisy evaluations.

Note that optimizing $\mathbb{E}\left[C_t(q_t)\right]$ with zeroth order information (its noisy evaluations) is an one dimensional continuous bandit problem (see Bubeck et al. (2009); Kleinberg (2005)). Standard bandit control methods with suitable discretization achieves regret $O\left(T^{\frac{2}{3}}\right)$. In the next two sections, we describe two variations of bandit problems and their corresponding learning algorithms. In the last section, we go back to the newsvendor problem with setup cost, applying our either of new learn algorithms to achieve the almost optimal regret rate — $O\left(\sqrt{\log(T)T}\right)$.

2.1 Online Learning with One-Side Zeroth Ordering Information and SAS Algorithm

In this section, we consider a discrete bandits control problem with so-called one-side feedback information. At each period t, an agent facing J actions (or bandit arms) selects one arm at every time step. With each arm $j \in [J]$.

In each period $t = 1,2\ldots$, there is an associated random cost function $C_t : [J] \to \mathbb{R}$ mapping a set of J actions to its associated cost. For notational convenience, we place argument as sup-script: $C_t^j := C_t(j)$. We assume that C_t are i.i.d. across each period. We use C to denote the time generic random functions C_t for $t = 1,2\ldots$, i.e., $C = C_t$ in distribution, and define the mean cost $\mu := \mathbb{E}[C]$. For each $j \in [J]$, we assume C^j if sub-gaussian with parameter σ:

$$\mathbb{E}e^{\lambda(C^j - \mu^j)} \le \frac{\sigma^2\lambda^2}{2} \text{ for all } \lambda \ge 0.$$

By Hoeffding bound, with probability as least $1-\delta$, we have

$$\left|\hat{\mu}_t^j - \mu^j\right| \le \sqrt{\frac{2\sigma^2 \log(2/\delta)}{t}} \tag{2.3}$$

where $\hat{\mu}_t^j := \frac{1}{t}\sum C_t^j$.

The distribution of C_t are unknown to the firm. At each period t, the firm chooses an action $j_t \in [J]$, and observe the one-side feedback $\left\{C_t^j : \text{ for } j = [j_t]\right\}$. The goal is to minimize the *regret*:

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{N}\left(C_t^{j_t} - C_t^{j^*}\right)\right], \tag{2.4}$$

where $j^* := \underset{j\in[J]}{\arg\min}\mu^j$.

Algorithm 2.1 Shrinking Active Set algorithm (SAS)

```
Parameters: confidence levels $\{\Delta_t\}_{t=1}^{T}$
Initialize active set $\mathcal{A}_1 = \{1, \ldots, J\}$
For $t = 1, 2, \ldots$, play maximum active action $j_t = \max\{\mathcal{A}_t\}$ and update active set
    $\mathcal{A}_{t+1} = \left\{ j \in \mathcal{A}_t : \hat{\mu}_t^j - \min_{k \in \mathcal{A}_t} \hat{\mu}_t^k \leq \Delta_t \right\}.$    (2.5)
```

Theorem 2.1. *Playing SAS algorithm with parameters* $\Delta_t = 2\sqrt{\frac{2\sigma^2 \log(2JT^2)}{t}}$*., we achieve regret*

$$\mathcal{R}_T = O\Big(\sqrt{\log(JT)T}\Big).$$

Proof. Define events

$$A := \left\{ \text{ For all } t \in [T], j \in [J] \text{ we have } \left|\hat{\mu}_t^j - \mu^j\right| \leq \Delta_t/2 \right\}.$$

$$A^c := \left\{ \text{ There exsits } t \in [T], j \in [J] \text{ such that } \left|\hat{\mu}_t^j - \mu^j\right| > \Delta_t/2 \right\}.$$

By Hoeffding bound (2.3) and union bound, we have

$$\mathbb{P}[A^c] \leq TJ\frac{1}{JT^2} = \frac{1}{T}.$$

Given event A holds, for any $j \in [J]$ and $t \in [T]$, we have

$$\hat{\mu}_t^{j^*} - \hat{\mu}_t^j \leq \hat{\mu}_t^{j^*} - \mu^{j^*} + \mu^j - \hat{\mu}_t^j \leq \Delta_t,$$

which implies

$$\hat{\mu}_t^{j^*} - \min_j \hat{\mu}_t^j \leq \Delta_t.$$

Then comparing with the updating rule (2.5) on our active set in algorithm 2.1, we conclude that j^* will always remain in the active set of every iteration and never leave. Since $j_t \in A_t$, which implies j_t is not “removed” from the active set in the $(n-1)^{th}$ iteration, we have

$$\hat{\mu}_{t-1}^{j_t} - \hat{\mu}_{t-1}^{j^*} \leq \hat{\mu}_{t-1}^{j_t} - \min_{j \in \mathcal{A}_{t-1}} \hat{\mu}_{t-1}^j \leq \Delta_{t-1},$$

where the second inequality follows from our rule (2.5). Therefore, conditional on the event A,

$$\mu^{j_t} - \mu^* = \mu^{j_t} - \hat{\mu}_{t-1}^{j_t} + \hat{\mu}_{t-1}^{j_t} - \hat{\mu}_{t-1}^{j^*} + \hat{\mu}_{t-1}^{j^*} - \mu^* \leq \Delta_{t-1}/2 + \Delta_{t-1} + \Delta_{t-1}/2 = 2\Delta_{t-1}$$

with $\Delta_0 := 1$.

Thus we have,

$$\begin{aligned}\mathcal{R}_T &= \mathbb{P}[A]\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t} - \mu^{j^*}\right)|A\right] + \mathbb{P}[A^c]\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t} - \mu^{j^*}\right)|A^c\right] \\ &\leq \sum_{t=1}^{T}\Delta_{t-1} + 1 = O\left(\sqrt{\log(JT)T}\right)\end{aligned}$$

where the last equality holds by plugging in $\Delta_t = 2\sqrt{\frac{2\sigma^2\log(2JT^2)}{t}}$. □

In one related work (Alon et al. (2015)), the authors consider a more general setting with information feedback modeled by a graph. Their more general EXP3.G algorithm can be applied here. An obvious advantage of SAS algorithm is the better regret rate $O\left(\sqrt{\log(JT)T}\right)$ compared with their $O\left(\log(JT)\sqrt{T}\right)$. More importantly, in supply chain optimization, policy transitions is key issue for achieve optimal regret (Chen et al. (2018c); Shi et al. (2016)). Consider the applying rule $j_t = \max\{\mathcal{A}_t\}$ and updating rule (2.5), we have at most J policy transitions. In practice $J << T$, which makes SAS algorithm easy to be adapted into a supply chain optimization problem (ref. my paper).

One disadvantage of above SAS algorithm is that, to achieve $O\left(\sqrt{\log(JT)T}\right)$ regret rate, we need to tune hyper-parameters $\{\Delta_t\}_{t=1}^{T}$ according to time horizon T. But, practically, T is rarely known in advance by the agent. In the following theorem, we apply the so-called "doubling trick" to remove our policy's dependence on knowing time horizon T in advance.

Theorem 2.2. *(Doubling trick) We divide time periods into groups; For $m = 0, 1, \ldots,$ the m^{th} group contains 2^m periods: $\left\{2^m, \ldots, 2^{m+1} - 1\right\}$. For each group m, we start a new SAS policy with parameters $\{\Delta_t\}_{t=2^m}^{2^{m+1}-1}$ such that $\Delta_{2^m+s-1} = 2\sqrt{\frac{2\sigma^2\log(2J2^{m+1})}{s}}$. Then*

$$\mathcal{R}_T = O\left(\log(JT)\sqrt{T}\right).$$

Proof. By Theorem 2.1, given time horizon T, the SAS algorithm achieves regret $O\left(\sqrt{\log(JT)T}\right)$. More precisely, there exist some positive constant α_1 and α_2 such that

$$\mathcal{R}_T \leq \alpha_1\sqrt{\log(JT)T} + \alpha_2.$$

Now, suppose T is unknown. Consider the first $M+1$ groups ($m = 0, \ldots, M$) contains $1 + 2 +$

$2^2 + \cdots + 2^M = 2^{M+1} - 1$ periods in total. Let $M = \min\left\{m : 2^{m+1} - 1 \geq T\right\}$, and then

$$\begin{aligned}\mathcal{R}_T &\leq \sum_{m=0}^{M} \left\{ \sum_{t=2^m}^{2^{m+1}-1} (\mu_{I_t} - \mu^*) \right\} \leq \sum_{m=0}^{M} \alpha_1 \sqrt{\log(J2^m)\,2^m} + \alpha_2 \\ &\leq \alpha_1 \sqrt{\sum_{m=0}^{M} \log(J2^m)\,2^m} \cdot \sqrt{M+1} + \alpha_2 (M+1) \end{aligned} \tag{2.6}$$

where the second inequality is because, in group m, we applied SAS policy with horizon length 2^m, and the last inequality is from Holder's inequality.

Consider

$$\begin{aligned}\sum_{m=0}^{M} \log(J2^m)\,2^m &\leq \int_0^{M+1} \log(J2^m)\,2^m \\ &= \frac{2^{M+1}\left((M+1)\log 2 + \log J - 1\right) - \log J + 1}{\log 2} = O\left((M + \log J)\,2^{M+1}\right)\end{aligned}$$

Plugging into (2.6), we have

$$\mathcal{R}_T = O\left(\sqrt{M(M + \log J)\,2^{M+1}}\right) = O\left(\sqrt{\log T \log(JT)\,T}\right)$$

where the last equality is due to $\log_2 T \geq M$ (by definition of M). □

2.2 Bandits on Convex Functions

2.2.1 Introduction

In this section we consider a variation of multi-armed bandit problem where each arm corresponds to a function. We consider J sequences of i.i.d. random convex cost functions $\left\{C_t^j\right\}_{t=1}^{\infty}$ for $j \in [J]$. We use time generic notation C^j to denote C_t^j, and $\mu^j := \mathbb{E}\left[C^j\right]$. For each $t = 1, 2, \ldots$, the agent faces a two step decision: choose a function index j_t , and, then choose an action $x_t \in \mathcal{K}^{j_t}$, which leads to the cost $C^{j_t}(x_t)$. .

Assumption 2. *We make the following assumptions:*

1. *We assume each realization of C^j is convex with a bounded domain $\mathcal{K}^j$ has diameter at most β. i.e.,*

$$\sup_{x_1, x_2 \in \mathcal{K}^j} \|x_1 - x_2\| \leq \beta.$$

2. *For each $j \in [J]$, and $x \in \mathcal{K}^j$, for random cost $C^j(x)$ is subgaussian with parameter σ:*

$$\mathbb{E}\left[e^{\lambda C(a)}\right] \le e^{\frac{\sigma^2\lambda^2}{2}} \textit{ for all } \lambda > 0,$$

and, the random gradient has bounded second moment: $\mathbb{E}\left[\left\|\nabla C^j(x)\right\|^2\right] \le \xi^2$.

We consider the distribution of C^j for $j \in [J]$ is unknown to the agent, but, at each period, he can observe cost $C^{j_t}(x_t)$ and $\nabla C^{j_t}(x_t)$. At period t, the agent selected j_t, x_t based on previously observed information $\left\{C^{j_{t'}}(x_{t'}), \nabla C^{j_{t'}}(x_{t'}) \text{ for } t' < t\right\}$ to minimize *regret:*

$$\mathcal{R}_T = \sum_{t=1}^{T} \mathbb{E}\left[C^{j_t}(x_t) - C^*(x^*)\right]$$

where

$$C^*(x^*) := \min_{j\in[J], x\in\mathcal{K}^j} \mathbb{E}\left[C^j(x)\right].$$

2.2.2 SGD Lower Confidence Bound Algorithm

Note that the best function j^* minimizes function minimum value: $j^* = \arg\min_j \left(\min_{x\in\mathcal{K}^j} \mu^j(x)\right)$. The key to choose best function is to estimate function minimum values. We use the running average along the stochastic gradient path as the estimator:

$$\hat{\mu}_t^j := \frac{1}{t}\sum_{s=1}^{t} C_s^j(x_s).$$

In the following proposition , we derive a high probability bound for the lower confidence bound.

Proposition 2.1. *Fix index j. Choosing steps size $\eta_t = \frac{\beta}{\xi\sqrt{t}}$, we have, with probability at least* $1-\delta$

$$\mu_*^j \ge \hat{\mu}_t^j - \frac{\sqrt{2\sigma^2 \log(1/\delta)} + 1.5\xi\beta}{\sqrt{t}}$$

where $\mu_*^j := \min_{x\in\mathcal{K}^j} \mu^j(x)$.

Proof. Introducing a bridging term $\mathbb{E}\left[\hat{\mu}_t^j\right]$, we have

$$\hat{\mu}_t^j - \mu_*^j = \left(\hat{\mu}_t^j - \mathbb{E}\left[\hat{\mu}_t^j\right]\right) + \left(\mathbb{E}\left[\hat{\mu}_t^j\right] - \mu_*^j\right).$$

By Theorem 3.4 in (Hazan et al. (2016)), we have

$$\mathbb{E}\left[\hat{\mu}_t^j\right] - \mu_*^j \leq \frac{1.5\xi\beta}{\sqrt{t}}.$$

By Azuma's inequality,

$$\hat{\mu}_t^j - \mathbb{E}\left[\hat{\mu}_t^j\right] \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{t}}$$

with probability $1-\delta$. □

We denote by T_t^j the (random) number of times function j is selected up to period t.

Algorithm 2.2 SGD lower confidence bound algorithm (SLCB)

Parameters:

For each arm $j \in [J]$, define $\hat{\mu}_i^j$ as the running average of i observed costs for function j. Define the index

$$B_{t,i}^j = \begin{cases} \hat{\mu}_i^j - \frac{\sqrt{2\sigma^2 \log(t^4)} + 1.5\xi\beta}{\sqrt{i}} & \text{for } s \geq 1; \\ -\infty & \text{for } s = 0. \end{cases}$$

At time t, select function j_t that minimizing $B_{t,T_j(t-1)}^j$, play $x_t = x_{T_t^j - 1}^j$. Observe $C(x_t)$ and $\nabla C(x_t)$. Update

$$x_{T_t^j}^j = \mathbf{Proj}_{\mathcal{K}^j}\left(x_t - \nabla C(x_t)\right)$$

Proposition 2.2. *The regret of the MLCB policy satisfies*

$$\mathcal{R}_T = O\left(\sqrt{\log(T)JT}\right).$$

Proof. We first bound the expected numb of pulls for a suboptimal arms. More precisely, in the first two steps of the proof we prove that, for any j such that $\Delta_j > 0$,

$$\mathbb{E}\left[T_j(T)\right] \leq \left(\frac{2\sqrt{2\sigma^2 \log(T^4)} + 3\xi\beta}{\Delta_j}\right)^2 + 5. \tag{2.7}$$

First step.

We show that if $j_t = j$, then one of the following three inequalities is true:

$$B_{t,T_{t-1}^{j^*}}^{j^*} \geq \mu_*^* \tag{2.8}$$

$$\hat{\mu}^j_{T^j_{t-1}} < \mu^j_* - \frac{\sqrt{2\sigma^2 \log(t^4)} + 1.5\xi\beta}{\sqrt{T^j_{t-1}}} \tag{2.9}$$

$$T^j_{t-1} < \left(\frac{2\sqrt{2\sigma^2 \log T^4} + 3\xi\beta}{\Delta_j} \right)^2 . \tag{2.10}$$

Indeed, assume that all three inequalities are false. Then we have

$$\begin{aligned} B_{j^*, T_{j^*}(i-1), t} < \mu^*_* = \mu^j_* - \Delta_j &\le \mu^j_* - \frac{2\sqrt{2\sigma^2 \log t^4} + 3\xi\beta}{\sqrt{T^j_{t-1}}} \\ &\le \hat{\mu}^j_{T^j_{t-1}} - \frac{\sqrt{2\sigma^2 \log t^4} + 1.5\xi\beta}{\sqrt{T^j_{t-1}}} = B_j^{T_j(i-1), i} \end{aligned}$$

which implies, in particular, that $j_t \neq j$.

Second step.

Here we bound the probability that (2.8) or (2.9) hold.

$$\mathbb{P}\left[(2.8) \text{ or } (2.9) \text{ is true}\right] \le 2 \sum_{s=1}^{t} \frac{1}{s^4} \le \frac{2}{t^3}.$$

Now, let

$$u = \left\lceil \left(\frac{2\sqrt{2\sigma^2 \log T^4} + 3\xi\beta}{\Delta_j} \right)^2 \right\rceil .$$

Using the first step, we have

$$\begin{aligned} \mathbb{E}\left[T_j(T)\right] = \mathbb{E} \sum_{t=1}^{T} \mathbf{1}_{j_t = j} &\le u + \mathbb{E} \sum_{t=u+1}^{T} \mathbf{1}_{\{j_t = j \text{ and (2.10) is false }\}} \\ &\le u + \mathbb{E} \sum_{t=u+1}^{T} \mathbf{1}_{\{j_t = j \text{ and (2.8 or 2.9) is true }\}} \le u + \sum_{t=u+1}^{T} \frac{2}{t^3} \le u + 4 \end{aligned}$$

This concludes the proof of (2.7).

Third Step.

We consider a bridging algorithm, which, at period t, selects the same function j_t and play the optimal action x^j_* for function μ^j.

We see that

$$\begin{aligned}\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_t)-\mu^*_*\right)\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_t)-\mu^{j_t}(x_*)\right)\right]+\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_*)-\mu^*_*\right)\right]\end{aligned}\tag{2.11}$$

Consider the first term,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_t)-\mu^{j_t}(x_*)\right)\right] &= \mathbb{E}\left[\sum_{j=1}^{J}\sum_{t=1}^{T_T^j}\left(\mu^{j}(x_t)-\mu^{j_t}(x_*)\right)\right]\\ &\leq \mathbb{E}\left[\sum_{j=1}^{J}1.5\xi\beta\sqrt{T_T^j}\right] \leq 1.5\xi\beta\sqrt{\sum_{j=1}^{J}\mathbb{E}\left[T_T^j\right]}\sqrt{\sum_{j=1}^{J}1} = 1.5\xi\beta\sqrt{TJ}\end{aligned}\tag{2.12}$$

where the first inequality is due to online convex optimization theory (ref), and the second inequality is by applying Holder's inequality.

For the second term, we consider

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_*)-\mu^*_*\right)\right] = \sum_{j=1}^{J}\Delta_j\mathbb{E}\left[T_j(T)\right] \leq \sum_{j:\Delta_j>0}\left(\left(\frac{\sqrt{2\sigma^2\log T^4}+3\xi\beta}{\Delta_j}\right)^2+5\Delta_j\right)$$

where we plug in (2.7) for the second inequality. For $T \geq t_0$ with some n_0 such that

$$\min_{j:\Delta_j>0}\left(\frac{2v\sqrt{4\log n_0}+3\xi\sqrt{d}}{\Delta_j}\right)^2 \geq 5$$

and $2v\sqrt{4\log n_0} \geq 3\xi\sqrt{d}$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}(x_*)-\mu^*_*\right)\right] &= \sum_{j:\Delta^j>0}\Delta^j\mathbb{E}\left[T_T^j\right] = \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T_T^j}\sqrt{\mathbb{E}T_T^j}\\ &\leq \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T_T^j}\sqrt{\left(\frac{\sqrt{2\sigma^2\log T^4}+3\xi\beta}{\Delta^j}\right)^2+5\Delta^j}\\ &\leq \sum_{j:\Delta^j>0}\Delta^j\sqrt{\mathbb{E}T_T^j}\sqrt{2\left(\frac{2\sqrt{2\sigma^2\log T^4}}{\Delta^j}\right)^2} = 4\sum_{j:\Delta^j>0}\sqrt{\mathbb{E}T_T^j}\sqrt{2\sigma^2\log T^4}\end{aligned}$$

$$
\begin{aligned}
&\le 4\sqrt{2\sigma^2 \log T^4}\sqrt{\sum_{j:\Delta_j>0} \mathbb{E}T_j(n)}\sqrt{\sum_{j:\Delta_j>0} 1} = 4\sqrt{2\sigma^2 \log T^4}\sqrt{TJ} \\
&= O\left(\sqrt{\log(T)JT}\right). \qquad (2.13)
\end{aligned}
$$

where the first inequality is due to (2.7), the second inequality is due to $T \ge n_0$ and the last inequality is due to Holder's inequality. Thus, combining (2.11) (2.12) and (2.13), we have finished the proof. □

2.3 Application with Newsvendor Problem with Fixed Cost

2.3.1 One Side Information Approach

Recall the Newsvendor with Fixed cost problem. With ordering quantity q_t, we suffer cost

$$
\tilde{C}_t(q_t) := K\mathbf{1}_{\{q_t \ge Q\}} + cq_t - p\min(q_t, D_t).
$$

and observe sale information $\min(q_{t,}D_t)$. Note this observation gives us so-called side zeroth order information: We can evaluation $\tilde{C}_t(q)$ for all $q \le q_t$. To apply the Shrinking Active Set algorithm (algorithm 2.1), we first choose a discrete grid $\mathcal{J} := \left\{q^j := \frac{j\beta}{J} : j = 0, 1, 2, \dots, J\right\}$ of the ordering domain $[0,\beta]$. Then, the regret

$$
\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t(q_t) - \tilde{C}_t\left(q^{j^*}\right)\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t\left(q^{j^*}\right) - \tilde{C}_t(q^*)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t(q_t) - \tilde{C}_t\left(q^{j^*}\right)\right)\right] + T\mathbb{E}\left[\tilde{C}\left(q^{j^*}\right) - \tilde{C}(q^*)\right]
\end{aligned}
$$

where j^* is the best on-grid ordering quantity: $j^* := \arg\min_{j \in \mathcal{J}} \mathbb{E}\left[\tilde{C}\left(q^j\right)\right]$. The first term is regret if we restrict to on-grid ordering, and the second term is the discretization loss.

Theorem 2.3. *Assume $\tilde{C}_t(q_t)$ is sub-gaussian with parameter σ. Choose $J \le \left\lceil\sqrt{N}\right\rceil$, and apply SAS algorithm on ordering $\left\{q^0, \dots, q^J\right\}$ with confidence level $\Delta_t = 2\sqrt{\frac{2\sigma^2\log(2JT^2)}{t}}$, we have*

$$
\mathcal{R}_T = O\left(\sqrt{\log(T)T}\right).
$$

Proof. To bound the discretization loss, we note that, although $\mathbb{E}\left[\tilde{C}(q)\right]$ is not a Lipschitz function for $q \in [0,\beta]$ (it is discontinuous at $q = Q$), it is piece-wise Lipschitz with Lipschitz constant p on piece of domain $[0, Q)$ and $[Q,\beta]$.

Case 1: $q^* \in [0, Q)$

Define $\tilde{q} := \underset{q^j \in \mathcal{J} \cap [0,Q)}{\arg\min} \left| q^j - q^* \right|$. i.e., $\tilde{q}$ is the closest on-grid point in $[0, Q]$. By the definition of grid $\mathcal{J}$, $|\tilde{q} - q| \leq \frac{\beta}{J}$, and, since on $[0, Q)$, $\mathbb{E}\left[\tilde{C}(q)\right]$ is p-Lipschitz,

$$\mathbb{E}\left[\tilde{C}(\tilde{q})\right] - \mathbb{E}\left[\tilde{C}(q^*)\right] \leq \frac{p\beta}{J}.$$

Therefore, the discretization loss

$$\mathbb{E}\left[\tilde{C}\left(q^{j^*}\right) - \tilde{C}(q^*)\right] \leq \mathbb{E}\left[\tilde{C}(\tilde{q})\right] - \mathbb{E}\left[\tilde{C}(q^*)\right] \leq \frac{p\beta}{J} \leq \frac{p\beta}{\sqrt{T}+1}. \tag{2.14}$$

Case 2: $q^* \in [Q, \beta]$

Apply the same argument, with $\tilde{q} := \underset{q^j \in \mathcal{J} \cap [Q,\beta]}{\arg\min} \left| q^j - q^* \right|$.

On the other hand, following from Theorem 2.1 the on-grid regret is

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\tilde{C}_t(q_t) - \tilde{C}_t\left(q^{j^*}\right)\right)\right] = O\left(\sqrt{\log(JT)T}\right) = O\left(\sqrt{\log(T)T}\right) \tag{2.15}$$

since $J = \left\lceil \sqrt{N} \right\rceil$.

Combining (2.14) and (2.15), we have $\mathcal{R}_T = O\left(\sqrt{\log(T)T}\right)$. □

2.3.2 Bandits on Convex Functions Approach

To apply SGD lower confidence bound algorithm, we make our choice of q_t with two-step decision: We first decide whether to order above Q, and, then, decide how much to order. By (2.1), $\tilde{C}(q)$ is convex on either domain $[0, Q)$ or domain $[Q,\beta]$, and, in both domains, a stochastic $\tilde{\nabla}$ is given by (2.2).

Assume both $\tilde{\nabla}(q)$ and $\tilde{C}(q)$ are sub-gaussian with parameter σ and ξ^2 bounds the second moments of $\tilde{\nabla}$, we apply SGD lower confidence bound algorithm on the decision whether to order above Q, which achieves regret $O\left(\sqrt{\log(T)T}\right)$.

2.4 Appendix

Definition 1. *A random variable X with mean μ is* ***sub-exponential*** *if there are non-negative parameters (ν, b) such that*

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\nu^2\lambda^2}{2}} \text{ for all } |\lambda| < \frac{1}{b}.$$

Proposition 2.3. *For zero-mean random variable X, the following statements are equivalent:*

1. $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}}$ for all $|\lambda| \leq 1/b$

2. There are positive constants c_1 and c_2 such that $\mathbb{P}[|X| > t] \leq c_1 e^{-c_2 t}$ for all $t > 0$.

Proposition 2.4. *Suppose $X^1, \ldots, X^n$ are centered (v, b) sub-exponential random variables. Then, $\frac{1}{n}\sum_{i=1}^n X^i$ is $\left(\frac{v}{\sqrt{n}}, b\right)$ sub-exponential. Consequently, for all $t \geq 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X^i - \mathbb{E}X\right| \geq t\right] \leq 2e^{-\min\left(\frac{t^2 n}{2v^2}, \frac{t}{2b}\right)}.$$

Equivalently,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X^i\right| \geq \max\left(\sqrt{\frac{2v^2\log(2/\delta)}{n}}, \frac{2b\log(2/\delta)}{n}\right)\right] \leq \delta.$$

t	period index
$\mathcal{J}$	action domain
T	period horizon
J	$\lvert\mathcal{J}\rvert$
j	$j \in \mathcal{J}$
C_j	$C\left(a_j\right)$
a_t	action at period t
T	period horizon
C_t	random cost function at time t, iid across t.
∇_t	gradient of C_t
n	cycle index
G_n	cycle cost
G_n^π	cycle cost with policy π
L_n^π	cycle length with policy π
G	time generic G_n
L	time generic L_n
$\hat{G}_n^j$	$\mathbb{E}[G]$
π_n	policy at cycle n
$\bar{L}$	$\mathbb{E}[L]$
v	$\hat{G}_n^j$
N	cycle horizon
σ	sub-gaussian parameters
(v, b)	sub-exponential parameters
$\mathcal{R}_T$	T period regret
$\mathcal{R}_N$	N cycle regret

Table 2.1: Notations 1

CHAPTER 3

Data Driven Inventory Control with Lead Time

For any $x \in \mathbb{R}$, $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$.

3.1 Inventory Control with Lead Time

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

In the considered setting, any new order will stay in the pipeline for L periods before arrival. We thus use an L-dimensional vector to track inventory. In each period t, the starting inventory, or state of the system, is

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

where I_t is the on-hand inventory at the beginning of period t, and, for $t' = t - L + 1, \dots, t - 1$, $q_{t'}$ is the order placed in period t'. Clearly, all the entries of x_t are non-negative. For simplicity, let $q_t = 0$ for all $t \leq 0$.

In each period t, the sequence of events is as follows:

1. At the beginning of each period t, the firm observes the starting inventory x_t, and makes an ordering decision $q_t \geq 0$.

2. The demand D_t is realized. It is satisfied to the maximum extent by on-hand inventory I_t. Since demand is censored, the firm only observes sales quantity $\min(D_t, I_t)$. Thus, if $D_t \geq I_t$, the firm does not know the exact demand.

3. At the end of the period, each remaining on-hand inventory unit incurs a holding cost h, and each unsatisfied demand unit incurs a penalty cost p. As a result, the cost in period t, denoted by C_t, is
$$C_t = h(I_t - D_t)^+ + p(I_t - D_t)^- . \tag{3.1}$$

4. At last, the system proceeds to period $t+1$ with a system state x_{t+1} as
$$x_{t+1} = [I_{t+1}, q_{t-L+2}, \dots, q_t]$$
where the on-hand inventory is
$$I_{t+1} = q_{t-L+1} + (I_t - D_t)^+ . \tag{3.2}$$

3.1.1 Objective and Assumptions

Let $\mathcal{F}_t$ denote the historical sales information collected up to the beginning of period t, i.e.,
$$\mathcal{F}_t = \sigma(\min(D_{t'}, I_{t'}) \text{ for } t' = 1, \dots, t-1) \tag{3.3}$$
and $\mathcal{F}_0 := \{\mathbb{R}, \emptyset\}$. Let $\mathcal{X} \subseteq \mathbb{R}^L$ denote the space of state, i.e. $x_t \in \mathcal{X}$ for each t.

Definition 2. *A* policy $\pi : \mathcal{X} \to \mathbb{R}$ *is a mapping from state x to ordering decision q. A* learning algorithm *is a sequence of random polices* $\{\pi_t\}_{t=1}^{\infty}$ *such that π_t is $\mathcal{F}_t$-measurable.*

Note that states x_t and decisions q_t depend on π, but we make the dependency implicit for notation simplicity. Only when necessary, we use x_t^π and q_t^π to represent state and ordering decision of policy π in period t. Our goal is to find a policy π that minimizes the *long-run average expected cost*
$$v^\pi := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} C_t^\pi\right] \tag{3.4}$$
where C_t^π is the cost at time t if we apply policy π. However, even when the demand distribution is known, finding the optimal policy is intractable due to the curse of dimensionality. In this paper, we follow Huh et al. (2009) to use the best base-stock policy as the benchmark. The class of base-stock policies is parameterized by a single parameter S, and the ordering quantity in period t is
$$q_t = (S - \|x\|_1)^+ .$$
Note that
$$\|x_t\|_1 = I_t + \sum_{t'=t-L+1}^{t-1} q_i$$

is the inventory position at the beginning of period t. Thus, the base-stock policy orders to raise the inventory position to S if the starting inventory position is less than S, and orders nothing otherwise. We call S the *inventory target*, and the corresponding base-stock policy the S-base-stock policy . We refer to Huh et al. (2009) for the asymptotic optimality and the effectiveness of base-stock policies. We measure the performance of a learning algorithm $\{\pi_t\}_{t=1}^{\infty}$ by its *regret*

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T}\left(C_t^{\pi_t} - v^*\right)\right],$$

where $v^* = \min_{\pi\in\{\text{base-stock polies}\}} v^{\pi}$ is the optimal long-run average expected cost among all base-stock policies.

For a fixed base-stock policy π, the states $\left\{x_t^{\pi}\right\}_{t=1}^{\infty}$ form a Markov chain. Let $P^{\pi,m}(x,\cdot)$ denote the m-step transition probability given starting with x. i.e.,

$$P^{\pi,m}(x,A) = \mathbb{P}[x_{t+m} \in A | x_t = x]$$

for any $t = 1,2,\ldots$ and any measurable event $A \subseteq \mathcal{X}$.

Assumption 3. *Throughout this paper, we make the following assumptions.*

1. *(*Bounded range*) The manager has an a priori knowledge of an upper bound $\bar{S}$ on S^*, i.e., $0 \le S^* \le \bar{S}$.*

2. *(Lost-sales time) Let $L(x,S)$ denote the first lost-sales period, given starting state x and base-stock target S. We assume*
$$\mathbb{E}[L(x,S)] \le M$$
for any $x \in \mathcal{X}$ and $S \le \bar{S}$.

3. *(*Uniform ergodic*) There exits a probability measure φ on $\mathcal{X}$, $\lambda > 0$, and an integer $m \ge 1$ such that for any base-stock policy π, we have*
$$P^{\pi,m}(x,\cdot) \ge \lambda\varphi(\cdot)$$
for each $x \in \mathcal{X}$.

These assumptions are mild. A base-stock policy without any lost-sales is clearly sub-optimal. To exclude the case explicitly, we assume a constant M bounding the expected first lost-sales time for all considered policies. For the uniform ergodic assumption, according to Huh and Rusmevichientong (2009), $\mathbb{P}\left[D \le \frac{S}{L+1}\right] > 0$ is its sufficient condition. By Theorem 3.2 in Appendix, the uniform ergodic condition is equivalent to $\sup_x \left\|P^{\pi,t}(x,\cdot)-\nu\right\|_1 \le \gamma\alpha^t$ for some $\gamma > 0$ and $\alpha \in (0,1)$. In this paper, we treat $\bar{S}$, M, m, λ, γ, α as constants.

Our problem has the same setting as the problem considered in Huh et al. (2009); Zhang et al. (2019). In their papers, stochastic gradient descent (SGD) plays the core role in their algorithm. However, SGD is generally not suitable for discrete product. In paper Huh and Rusmevichientong (2009), to apply SGD, the authors assume a so-called "lost-sale indicator condition", which cannot be used in practice. In this paper, we indeed allow both continuous ($D_t, q_t \in \mathbb{R}^+$) and discrete setting ($D_t, q_t \in \mathbb{Z}^+$), and we design a learning algorithm that achieves almost optimal regret rate $O\left(\sqrt{\log(T)T}\right)$. Even if restricted to continuous setting, the algorithm proposed in Zhang et al. (2019) achieves $O\left(\sqrt{T}\right)$ regret, which is just asymptotically better than our $O\left(\sqrt{\log(T)T}\right)$ regret by a $\sqrt{\log(T)}$ factor. However, their regret depends on L exponentially. Our constant depends on L linearly, which makes our learning algorithm more appealing when L is large.

3.2 Parallel Evaluation Algorithm

3.2.1 Pseudo-Cost

Instead of using first order method (e.g. SGD), we design an algorithm that uses zeroth order method: In each period, we select a policy based on its empirical performance. However, as in (3.1), due to the censored demand, when lost sales occur at a period t, the cost C_t is not fully observable. We seek an alternative measurement of that cost that is observable to the firm. Note that the censored part of the cost C_t is $p(I_t - D_t)^-$, which is not observable whenever a lost sale occurs. Nevertheless, we can decompose it as

$$p(I_t - D_t)^- = pD_t - p\min(I_t, D_t).$$

The simple transformation above is crucial for our analysis because (a) the first term pD_t is independent of any feasible policy, and (b) the second term $p\min(I_t, D_t)$ is policy dependent but observable. Thus, we define what-we-call inventory *pseudo cost* by dropping the first term

$$\tilde{C}_t := C_t - pD_t = h(I_t - D_t)^+ - p\min(I_t, D_t). \tag{3.5}$$

By the definition of long-run average cost in (3.4), we see

$$\nu^\pi = \mu^\pi + p\mathbb{E}[D].$$

where $\mu^\pi := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \tilde{C}_t^\pi\right]$. Regret can be formulated using pseudo cost as below.

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t^{\pi_t} - \tilde{\mu}^*\right)\right].$$

3.2.2 Parallel Evaluation Algorithm

The notion of pseudo cost $\tilde{C}_t$ enables us to evaluate a policy. To make the evaluation more efficient, we design an algorithm that parallelizes the process, which is based on the following proposition.

Proposition 3.1. *Given two sequences of inventory targets S^1 and S^2, let x_t^1 (or x_t^2) denote the inventory state at period t for applying base-stock policy with inventory target S^1 (or S^2). Then, we have $x_t^1 \geq x_t^2$ componentwise for all t.*

Proof. Since state x_t consists of on-hand inventory I_t and undelivered orders $[q_{t-L+1}, \ldots, q_{t-1}]$. To show $x_t^1 \geq x_t^2$, it suffices to show $I_t^1 \geq I_t^2$ and $q_t^1 \geq q_t^2$. Also, since we are running a base-stock policy, except for the first period, we have $q_t = \min(D_{t-1}, I_{t-1})$. i.e., the order quantity is equal to the sales quantity of the previous period. Thus, we only need to show $I_t^1 \geq I_t^2$.

We prove by induction on t. For $t = 1, \ldots, L$, $I_t^1 = I_t^2 = 0$. Suppose at periods $t = 1, \ldots, t'$ for some $t' > L$, we have $I_t^1 \geq I_t^2$. At period $t+1$, by (3.2), we have

$$I_{t+1} = q_{t-L+1} + (I_t - d_t)^+ .$$

Note that since we are running base-stock policy, $q_{t-L+1} = \min(D_{t-L}, I_{t-L})$. By induction, $I_{t-L}^1 \geq I_{t-L}^2$, so we $q_{t-L+1}^1 \geq q_{t-L+1}^2$. Then, $I_{t+1}^1 \geq I_{t+1}^2$. □

Thus, given two base-stock policies π^1 with targets S^1 and π^2 with target S^2, if $S^1 \geq S^2$, we can use the sales information $\min\left(I_t^1, D_t\right)$ collected from applying policy π^1 to evaluate the sales $\min\left(I_t^2, D_t\right)$ from applying policy π^2. Let $\hat{C}_t^\pi$ denote the empirical average cost up to period t by applying policy π. i.e.,

$$\hat{C}_t^\pi := \frac{1}{t}\sum_{t'=1}^{t} \tilde{C}_{t'}^\pi.$$

Based on an ergodic theory argument, we see that $\hat{C}_t^\pi$ is a good approximation of μ^π.

Proposition 3.2. *With probability at least $1-\delta$, we have*

$$\left|\hat{C}_t^\pi - \mu^\pi\right| \leq \frac{m \max(h,p)\overline{S}}{\lambda}\left(\frac{2}{t} + \frac{\sqrt{2\log(2/\delta)}}{\sqrt{t}}\right)$$

Proof. By (3.5), $\left|\tilde{C}_t\right| \le \max(h,p)\overline{S}$. By *the uniform ergodic in Assumption 3,* Theorem 3.3 in appendix implies that

$$\mathbb{P}\left[\left|\hat{C}_t^{\pi} - \mu^{\pi}\right| \ge \epsilon\right] \le 2\exp\left(-\frac{(\alpha\epsilon t - 2)^2}{2t}\right).$$

where $\alpha = \frac{\lambda}{m\max(h,p)\overline{S}}$. □

i.e., with probability at least $1-\delta$, we have

$$\left|\hat{C}_t^{\pi} - \mu^{\pi}\right| \le \frac{m\max(h,p)\overline{S}}{\lambda}\left(\frac{2}{t} + \frac{\sqrt{2\log(2/\delta)}}{\sqrt{t}}\right).$$

With this we are able to design a parallel evaluation algorithm as follows.

Parameters. Let $\mathcal{S} = \{S_1, \ldots, S_J\}$ be the set of increasing ordering targets, and Δ_t be the so-called confidence bound of the t^{th} period. We will specify how to choose the optimal values of these parameters later in our algorithm analysis.

Initialization.

1. The algorithm maintains a active set $\mathcal{A}_t$ that contains the favorable candidates of optimal solution after the $t-1$ periods. We initialize $\mathcal{A}_1 = \{1, \ldots, J\}$. As the learning algorithm proceeds and t increases, the set $\mathcal{A}_t$ decreases. (We are removing the unlikely candidates gradually when information is sufficient.)

2. For each $j \in \mathcal{A}_t$, we keep track of the so-called virtual state x_t^j, which is equal to the state at period t assuming we run base-stock policy with inventory target S^j from the beginning. We initialize x_1^j as L dimensional zero vector for each $j \in \mathcal{A}_1$.

3. For each $j \in \mathcal{A}_t$, we also keep track of the average pseudo cost $\hat{C}_t^j$. We initialize $\hat{C}_0^j = 0$ for each $j \in \mathcal{A}_1$.

Main Loop.

1. At each period t, we first find out the maximum active index j_t. i.e., $j_t \in \max(\mathcal{A}_t)$. We apply base-stock policy with target $S_t = S^{j_t}$ with respect to virtual state $x_t^{j_t}$. i.e., we order

$$q_t = S_t^{j_t} - \left\|x_t^{j_t}\right\|_1.$$

With this, we achieve two benefits.

(a) (Information collection) By ordering with respect to the maximum inventory target S^{j_t} with respect to virtual state $x_t^{j_t}$, we can collect enough information to *simulate* all the other active policies. We are able to evaluate all active policies in parallel, and update the corresponding virtual states, which speed up our learning process.

(b) (Fast state transition) The real inventory state x_t is affected by all pervious inventory target $S_{t'}$ for $t' < t$, which make it hard to be analyzed. On the other hand, $x_t^{j_t}$ is the state assuming we keep applying S^{j_t} as inventory from the very begin, and we have a good understanding of it due to ergodic theory. By ordering with respect to the virtual inventory $x_t^{j_t}$, we force x_t to become $x_t^{j_t}$ fast: If the inventory target keeps unchanged, after L period x_{t+L} and $x_{t+L}^{j_t}$ has the same undelivered orders, and to make them equal, we only need to wait for a lost-sale period, which according to Assumption 3, is bounded by M in expectation.

2. Since x_t is great than each visual state x_t^j for $j \in \mathcal{A}_t$. We can compute the virtual sale $\min\left(I_t^j, D_t\right)$ for each active j. Then, for each $j \in \mathcal{A}_t$ we update the virtual state

$$x_{t+1}^j = \left[I_{t+1}^j, q_{t-L+2}^j, \ldots, q_t^j\right]$$

where $I_{t+1}^j := I_t^j - \min\left(I_t^j, D_t\right) + q_{t-L+1}^j$, and update the empirical average cost

$$\hat{C}_t^j = \frac{(t-1)\hat{C}_t^j + \tilde{C}_t^j}{t},$$

where $\tilde{C}_t^j := h\left(I_t^j - D_t\right)^+ - p\min\left(I_t^j, D_t\right)$ is the j^{th} virtual pseudo cost.

3. Based on the empirical performance measured by $\hat{C}_t^j$, we prune actives as follows

$$\mathcal{A}_{t+1} = \left\{ j \in \mathcal{A}_t : \hat{C}_t^j - \min_{j'} \hat{C}_t^{j'} \leq \Delta_t \right\}.$$

i.e., we remove all polices who empirical average cost is great than the optimal empirical average cost by more that the confidence bound Δ_t.

This concludes the description of our parallel evaluation algorithm learning algorithm. For the convenience of practical implementation, we also provide a detailed pseudo code in Algorithm 3.1.

A pivotal step in the algorithm is the to simulate all the active policies. We shall discuss why the above specified rules can indeed collect sufficient demand information. Note to simulate a base-stock policy j, we need its sales information $\min\left(I_t^j, D_t\right)$. What we can observe is the real sales data $\min(I_t, D_t)$. Thus, to make the simulate, it suffices to have $I_t \geq I_t^j$ for each active $j \in \mathcal{A}_t$ in each period. Let τ_k be the be starting period of the k^{th} unique policy. Let's induction on k. In periods

$[\tau_1, \tau_2)$, we use S^J as target, and $x_t = x_t^J$. By Proposition 3.1, $x_t \geq x_t^j$ for any $j \in [J]$. Suppose up to period τ_k we have $x_t \geq x_t^j$ for all $j \in \mathcal{A}_t$. During periods $[\tau_k, \tau_{k+1})$, we apply base-stock policy with target $S^{j_{\tau_k}}$ using virtual state $x_t^{j_{\tau_k}}$, which makes $q_t = q_t^{j_t}$ and implies $x_t \geq x_t^{j_t}$ for $t = \tau_k + 1, \ldots, \tau_{k+1}$. Since $x_t^{j_t}$ is the state with respect to the highest active target at period t. Again, by Proposition 3.1, $x_t^{j_t} \geq x_t^j$ for $j \in \mathcal{A}_t$ for $t = \tau_k + 1, \ldots, \tau_{k+1}$. This completes the induction reasoning. In each period, we have $x_t^{j_t} \geq x_t^j$ for $j \in \mathcal{A}_t$. In particular, $I_t \geq I_t^j$ for each active $j \in \mathcal{A}_t$, and this simulation step is indeed valid. Since our algorithm focuses on a set of policy candidates $\mathcal{S}$, we first analyze that the regret with respect the base policy in $\mathcal{S}$. More precisely, for an learning algorithm π, we define

Algorithm 3.1 Parallel evaluation algorithm

1. Parameters: A set of candidate orders target $\mathcal{S} = \left\{S^1, \ldots, S^J\right\}$, and confident bounds $\{\Delta_t\}_{t=1}^T$.

2. Initialize active set $\mathcal{A}_1 = \{1, \ldots, J\}$. For each $j \in \mathcal{A}_1$, define j^{th} virtual state $x_1^j = \mathbf{0} \in \mathbb{R}^L$, and empirical average cost $\hat{C}_0^j = 0$.

3. At time t, select $j_t = \max(\mathcal{A}_t)$, make order $q_t = S^{j_t} - \left\|x_t^{j_t}\right\|$ (apply base-stock policy with target $S_t = S^{j_t}$ with virtual state $x_t^{j_t}$)

   - Use the observed sales $\min(I_t, D_t)$ to simulate all the policies in $\mathcal{A}_t$: For all $j \in \mathcal{A}_t$, compute

     $$\begin{aligned} \tilde{C}_t^j &= h\left(I_t^j - D_t\right)^+ - p \min\left(I_t^j, D_t\right), \\ q_t^j &= S^j - \left\|x_t^j\right\|_1 \end{aligned}$$

     and update

     $$\begin{aligned} \hat{C}_t^j &= \frac{(t-1)\hat{C}_{t-1}^j + \tilde{C}_t^j}{t}, \\ x_{t+1}^j &= \left[I_t^j - \min\left(I_t^j, D_t\right) + q_{t-L+1}^j, q_{t-L+2}^j, \ldots, q_t^j\right]. \end{aligned}$$

   - Update active set

     $$\mathcal{A}_{t+1} = \left\{j \in \mathcal{A}_t : \hat{C}_t^j - \min_{j'} \hat{C}_t^{j'} \leq \Delta_t\right\}.$$

$$\mathcal{R}_T^{\mathcal{S}} = \mathbb{E}\left[\sum_{t=1}^T \left(C_t^\pi - \mu^{j^*}\right)\right]$$

where μ^j denote the long-run average cost for running base-stock policy with target S^j and $j^* :=$

$\arg\min_{j} \mu^j$.

Theorem 3.1. *Applying the parallel evaluation algorithm above with* $\Delta_t = 2\frac{m\max(h,p)\bar{S}}{\lambda}\left(\frac{2}{t} + \frac{\sqrt{2\log\left(2JT^{3/2}\right)}}{\sqrt{t}}\right)$, *we have*

$$\mathcal{R}_T^{\mathcal{S}} = O\left(\sqrt{\log(JT)T} + J\right).$$

Proof. Consider

$$\begin{aligned}\mathcal{R}_T^{\mathcal{S}} &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t - \mu^{j^*}\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t - \tilde{C}_t^{j_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t^{j_t} - \mu^{j_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t} - \mu^{j_*}\right)\right]\end{aligned}$$

Step 1. We will show $\mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t - \tilde{C}_t^{j_t}\right)\right] = O(J)$.

Let K denote the total number of unique policies apply for $t = 1, 2, \ldots, T$. Note $K \leq J$. For $k = 1, \ldots, K$, let j_k denote the k^{th} unique policy we applied and τ_k denote this first time we apply policy j_k. Consider

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t - \tilde{C}_t^{j_t}\right)\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1}\left(\tilde{C}_t - \tilde{C}_t^{j_t}\right)\right] \\ &= \sum_{k=1}^{J}\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1}\left(\tilde{C}_t - \tilde{C}_t^{j_t}\right)\right]\end{aligned}$$

where $\tau_k := T + 1$ for $k > K$.

During periods $[\tau_k, \tau_{k+1})$, we apply base-stock policy with inventory target S^{j_k} with respect to inventory state x_t^j. If we keep S fixed, after L periods, x_t and $x_t^{j_k}$ have the same the undelivered orders. Then, To make $x_t = x_t^{j_k}$, we only need wait until the first lost-sales, which, because of the lost-sale time assumption in Assumption 3, happens within M periods in expectation. Thus, we have

$$\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1}\left(\tilde{C}_t - \tilde{C}_t^{j_k}\right)\right] \leq \max(h,p)\bar{S}\,\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1}\mathbf{1}_{\left\{x_t \neq x_t^{j_k}\right\}}\right] \leq \max(h,p)\bar{S}\,(L+M),$$

where the first inequality is due to the one period pseudo cost is bounded by $\max(h,p)\bar{S}$.

Step 2. We shall show $\mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t^{j_t} - \mu^{j_t}\right)\right] = O(1)$.

Consider

$$\mathbb{E}\left[\tilde{C}_{t+1}^{j_t} - \mu^{j_t}\right] = \int\left(P^{j,t}(x_1, x) - \nu^j(x)\right)\tilde{C}(x)\,dx$$

$$\leq \left\|P^{j,t}(x_1,x) - \nu^j\right\|_1 \left\|\tilde{C}\right\|_\infty$$
$$\leq \max(p,h)\bar{S}\gamma\alpha^t$$

where the last inequality is dual to uniform ergodic assumption. Then,

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\tilde{C}_t^{j_t} - \mu^{j_t}\right)\right] \leq \max(p,h)\bar{S}\gamma\sum_{t=1}^{T}\alpha^{t-1} \leq \frac{\max(p,h)\bar{S}\gamma}{\alpha}.$$

Step 3. We will show $\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t} - \mu^{j_*}\right)\right] = O\left(\sqrt{\log(T)T}\right)$

Define events

$$A := \left\{ \text{For all } t \in [T], j \in [J] \text{ we have } \left|\hat{C}_t^\pi - \mu^\pi\right| \leq \Delta_t/2\right\}.$$
$$A^c := \left\{ \text{There exsits } t \in [T], j \in [J] \text{ such that } \left|\hat{C}_t^\pi - \mu^\pi\right| > \Delta_t/2\right\}.$$

By Proposition 3.2 and union bound, we have

$$\mathbb{P}[A^c] \leq TJ\frac{1}{JT^{3/2}} \leq \frac{1}{\sqrt{T}}.$$

Given event A holds, for any $j \in [J]$ and $t \in [T]$, we have

$$\hat{C}_t^{j^*} - \hat{C}_t^j \leq \hat{C}_t^{j^*} - \mu^{j^*} + \mu^j - \hat{C}_t^j \leq \Delta_t,$$

which implies

$$\hat{C}_t^{j^*} - \min_j \hat{C}_t^j \leq \Delta_t.$$

Then compared with the active set updating rule in algorithm 3.1, we conclude that j^* will always remain in the active set of every iteration and never leave. Since $j_t \in \mathcal{A}_t$, which implies j_t is not "removed" from the active set in the $(n-1)^{th}$ iteration, we have

$$\hat{C}_{t-1}^{j_t} - \hat{C}_{t-1}^{j^*} \leq \hat{C}_{t-1}^{j_t} - \min_{j \in \mathcal{A}_{t-1}} \hat{C}_{t-1}^j \leq \Delta_{t-1},$$

where the second inequality follows from our active set updating rule. Therefore, conditional on the event A,

$$\mu^{j_t} - \mu^* = \mu^{j_t} - \hat{C}_{t-1}^{j_t} + \hat{C}_{t-1}^{j_t} - \hat{C}_{t-1}^{j^*} + \hat{C}_{t-1}^{j^*} - \mu^* \leq \Delta_{t-1}/2 + \Delta_{t-1} + \Delta_{t-1}/2 = 2\Delta_{t-1}$$

with $\Delta_0 := 1$.

Thus we have,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}-\mu^{j_*}\right)\right] &= \mathbb{P}[A]\,\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}-\mu^{j^*}\right)|A\right]+\mathbb{P}[A^c]\,\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu^{j_t}-\mu^{j^*}\right)|A^c\right]\\ &\leq \sum_{t=1}^{T}\Delta_{t-1}+1=O\left(\sqrt{\log(JT)\,T}\right)\end{aligned}$$

where the last equality holds by plugging in $\Delta_t = 2\frac{m\max(h,p)\overline{S}}{\lambda}\left(\frac{2}{t}+\frac{\sqrt{2\log\left(2JT^{3/2}\right)}}{\sqrt{t}}\right)$. □

To bound regret $\mathcal{R}_T$, we need to choose the good candidate set $\mathcal{S}$. By the ergodic assumption, we have, for a fixed based-stock policy, the Markov chain $\{x_t\}_{t=1}^{\infty}$ converges to a limiting random vector x_∞. Let I_∞ denote the first component of x_∞, which is the limiting random on-hand inventory. Then long-run average pseudo cost can be written as

$$\mu = \mathbb{E}\left[h\left(I_\infty - D\right)^+ - p\min\left(I_\infty, D\right)\right]. \tag{3.6}$$

(see Huh and Rusmevichientong (2009)for detailed arguments.)

In continuous setting, since change target S by ϵ can at most change I_∞ by ϵ, (3.6) implies μ is $\max(h,p)$ Lipschitz. Thus, if we choose

$$\mathcal{S}=\left\{S^j := \frac{j}{J}\overline{S} \text{ for } J=1,\ldots,\lceil T\rceil\right\}.$$

Let S^* be the optimal ordering target. By the definition of $\mathcal{S}$, we have $\left|S^{j_0}-S^*\right| \leq \frac{\bar{S}}{\sqrt{T}}$, where $j_0 := \underset{j\in[J]}{\arg\min}\left|S^j - S^*\right|$. Therefore,

$$\mu^{j^*}-\mu^* \leq \mu^{j_0}-\mu^* \leq \bar{S}\max(h,p)\left|S^{j_0}-S^*\right| \leq \bar{S}\max(h,p)/\sqrt{T}.$$

and

$$\begin{aligned}\mathcal{R}_T &= \mathcal{R}_T^{\mathcal{S}} + T\left(\mu^{j^*}-\mu^*\right)\\ &= O\left(\sqrt{\log(T)\,T}+\sqrt{T}\right)+T\left(\bar{S}\max(h,p)/\sqrt{T}\right)\\ &= O\left(\sqrt{\log(T)\,T}\right).\end{aligned}$$

In discrete cast, we choose

$$\mathcal{S}=\left\{S^j := \left\lceil\frac{j}{J}\bar{S}\right\rceil \text{ for } J=1,\ldots,\lceil T\rceil\right\}.$$

If $\bar{S} \le \sqrt{T}$, we simply let $\mathcal{S} = \left\{1, 2, \ldots \bar{S}\right\}$, then

$$\mathcal{R}_T = \mathcal{R}_T^{\mathcal{S}} = O\left(\sqrt{\log(T)T} + \sqrt{T}\right) = O\left(\sqrt{\log(T)T}\right)$$

If $\bar{S} > \sqrt{T}$, define

$$\mathcal{S} = \left\{S^j := \left\lceil \frac{j}{J}\bar{S} \right\rceil \text{ for } J = 1, \ldots, \lceil T \rceil \right\}$$

As in the continuous case define $j_0 = \underset{j \in [J]}{\arg\min} \left|S^j - S^*\right|$, then $\left|S^{j_0} - S^*\right| \le \frac{\bar{S}}{\sqrt{T}} + 1$, where the “add 1” is due to rounding error. Since $\bar{S} > \sqrt{T}$, we have $\left|S^{j_0} - S^*\right| \le 2\frac{\bar{S}}{\sqrt{T}}$. Therefore,

$$\mu^{j^*} - \mu^* \le \mu^{j_0} - \mu^* \le \bar{S} \max(h, p) \left|S^{j_0} - S^*\right| \le \bar{S} \max(h, p) / \sqrt{T}.$$

and

$$\begin{aligned}
\mathcal{R}_T &= \mathcal{R}_T^{\mathcal{S}} + T\left(\mu^{j^*} - \mu^*\right) \\
&= O\left(\sqrt{\log(T)T} + \sqrt{T}\right) + T\left(2\bar{S} \max(h, p) / \sqrt{T}\right) \\
&= O\left(\sqrt{\log(T)T}\right).
\end{aligned}$$

Thus, we have produce a learning algorithm that achieves $O\left(\sqrt{\log(T)T}\right)$ regret rate in both continuous and discrete settings.

3.3 Conclusion

In this paper, we have proposed the a nonparametric learning algorithm for managing stochastic inventory systems with lead time under censored demand information, and showed that the regret is $O(\sqrt{\log(T)T})$, which is provably optimal up to a square root of a logarithmic factor. Compared with previous works, our algorithm can handle both discrete and continuous setting. Moreover, our regret bound depends linearly on lead time L. Even if restricted to continuous setting, when L is large, our learning policy is more appealing than previous algorithms whose regret depends exponentially on L.

3.4 Appendix

Theorem 3.2. *Let $P^t(x, \cdot)$ denote the The following are equivalent.*

1. $\sup_x \left\|P^t(x, \cdot) - \nu\right\|_1 \to 0.$

2. $\sup_x \left\| P^t(x,\cdot) - \nu \right\|_1 \le \gamma\alpha^t$ *for some* $\gamma > 0$ *and* $\alpha \in (0,1)$.

3. *There exists a probability measure* φ *on state space* $\mathcal{X}$, $\lambda > 0$ *and an integer* $m \ge 1$ *such that*

$$P^m(x,\cdot) \ge \lambda\varphi(\cdot) \textit{ for all } x \in \mathcal{X}.$$

Theorem 3.3. *(Hoeffding's inequality on Markov chain) Let* $X_1, X_2 \ldots$ *be a Markov chain. Suppose there exists a probability measure* φ *on state space* $\mathcal{X}$, $\lambda > 0$ *and an integer* $m \ge 1$ *such that*

$$P^m(x,\cdot) \ge \lambda\varphi(\cdot).$$

Then

$$\mathbb{P}\left[\left|\frac{\sum_{t=1}^n f(X_t)}{n} - \tilde{\mathbb{E}}[f(X)]\right| \ge \epsilon\right] \le 2\exp\left(\frac{(\alpha\epsilon n - 2)^2}{2n}\right)$$

where $\tilde{\mathbb{E}}$ *takes expectation with respect to the limiting distribution of the Markov chain and constant* $\alpha = \frac{\lambda}{m\|f\|_\infty}$.

t	period index
S	base-stock ordering target
T	time horizon
D_t	demand at period t
q_t	ordering quantity at period t
I_t	on-hand inventory at beginning of period t
x_t	inventory state at beginning of period t
y_t	inventory state after ordering at period t
h	unit holding cost
p	unit lost-sales penalty cost
C_t	cost at period period t
$\tilde{C}_t$	pseudo-cost at period t
v^π	expected long-run average cost of policy π
v^*	optimal v^π among all base-stock polices
μ^π	expected long-run average pseudo cost of policy π
μ^*	optimal μ^π among all base-stock polices
$\mathcal{R}_T$	cumulative regret
$\mathcal{S}$	set of inventory target candidates
J	size of $\mathcal{S}$
$\tilde{C}_t^j$	virtual pseudo-cost at period t of j^{th} policy
$\hat{C}_t^j$	t-period empirical average cost of j^{th} policy
j^*	index of optimal policy is $\mathcal{S}$
$\mathcal{A}_t$	active set at period t
$\bar{S}$	upper bound of inventory target
M	upper bound of expected lost-sale time
$m, \lambda, \gamma, \alpha$	constants related to uniform ergodic condition

Table 3.1: Notations 2

CHAPTER 4

Multi-Product Base Stock Inventory Control

Notation: For $a, b \in \mathbb{R}^n$, $a \cdot b \in \mathbb{R}$: inner production; $\max(a,b), \min(a,b), ab, a/b, \exp(a) \in \mathbb{R}^n$ componentwise max min, product, division and exponential; $[a]^+ = \max(a, \mathbf{0})$ and $[a]^- = \max(-a, \mathbf{0})$.

4.1 Multi-Product Stochastic Inventory System

We consider a stochastic T-period n-product inventory system under a warehouse-capacity constraint M. The firm has no knowledge of the true underlying demand distribution a priori, but can observe past sales data (i.e., censored demand data), and make adaptive inventory decisions based on the available information.

For each period $t = 1, ..., T$ and each product $j = 1, ...J$, we denote the demand of product j in period t by a random variable D_t^j . For notation simplicity, we use $D_t = (D_t^1, ..., D_t^J)$ to denote the random demand vector in period t.

Assumption 4. *We assume the following regularity conditions on demand.*

1. *The demand vector $D_t = (D_t^1, \dots, D_t^J)$ is i.i.d. over time $t = 1, 2, \dots$.*

2. *For each product j and for each period t, D_t^j is a continuous random variable defined on a finite support $[0, \beta]$, and $\mathbb{E}[D_t^j] \geq \alpha$ for some real number $\alpha > 0$.*

Let $\mathcal{F}_t$ denote the information collected up to the beginning of period t, which includes all realized demand and past decisions. A data-driven policy π is a sequence of functions $y_t = \pi_t(x_t, \mathcal{F}_t)$, $t = 1, ..., T$, mapping beginning inventory x_t and $\mathcal{F}_t$ (state) into ending inventory y_t (decision) while satisfying $y_t \geq x_t$ and the warehouse-capacity constraint. Note that when the demand distribution is known a prior, it suffices to consider policies of the form $y_t = \pi_t(x_t)$, due to the assumed cross-time independence of demands. Given a data-driven policy π, we describe the sequence of events below.

1. At the beginning of period t, the firm observes the starting inventory $x_t = (x_t^1, ..., x_t^J)$.

2. The firm decides to order $q_t = (q_t^1, ..., q_t^J) \geq 0$, and the ending inventory $y_t = x_t + q_t$, where $y_t = (y_t^1, ..., y_t^J)$. We assume instantaneous replenishment. The total inventory level is restricted by warehouse capacity, i.e.,

$$y_t \in \Gamma := \left\{ y \in \mathbb{R}_+^J : ||y||_1 \leq M \right\}.$$

3. The demand D_t is realized, which is satisfied to the maximum extent with on-hand inventory. Unsatisfied demand units are lost, and the firm only observes the sales quantity (or censored demand), i.e., $min(D_t^J, y_t^J)$ for each product J in period t. The state transition is

$$x_{t+1} = (x_t + q_t - D_t)^+ = (y_t - D_t))^+.$$

4. The total cost of production, overage and underage at the end of period t is thus $c \cdot q_t + h \cdot (y_t - D_t)^+ + p \cdot (y_t - D_t)^-$, where $c = (c_1, ..., c_J), h = (h_1, ..., h_J)$ and $p = (p_1, ..., p_J)$ are the unit cost of purchasing, holding and lost-sales penalty, respectively. We note that the cost minimization model with lost-sales assumes that $p \geq c$.

Assuming the salvage value of any left-over product at the end of planning horizon equals its production cost, the total expected cost incurred by π can be written as

$$\begin{aligned} \mathcal{C}(\pi) &= \mathbb{E}\left[\sum_{t=1}^{T} c \cdot (y_t - x_t) + h \cdot (y_t – D_t)^+ + p \cdot (y_t - D_t)^- \right] - \mathbb{E}[c. \cdot x] \\ &= -c \cdot x_1 + \mathbb{E}\left[\sum_{t=1}^{T} c \cdot y_t + h \cdot (y_t – D_t)^+ + p \cdot (y_t - D_t)^- \right]. \end{aligned}$$

If the underlying distribution D_t is given a priori, the stochastic inventory control problem specified above can be formulated using dynamic programming (see Beyer et al. (2001)) with state variables x_t, control variables y_t (with $x_t \leq y \in \Gamma$), random disturbances D_t , and state transition $x_{t+1} = (y_t - d_t)^+$. It turns out that this problem is in fact “myopically” solvable, which will be discussed next.

Clairvoyant optimal policy. We first characterize the clairvoyant optimal policy where the distribution of D_t is known a priori. We define

$$C(a) = C_t(a) = c \cdot a + h \cdot (a – D_t)^+ + p \cdot (a - D_t)^-. \tag{4.1}$$

Let $y^* = \arg\min_{a \in \Gamma} \mathbb{E}[C(a)]$.

Theorem 4.1. *Under Assumption 4, when the demand distribution is known a priori, ordering up* *to y^* in each period is optimal, with expected per-period cost* $\mathbb{E}[C(y^*)]$.

The proof of this theorem relies on verifying a sufficient condition known as substitute property provided by Ignall and Veinott Jr (1969).

We measures the performance of a data-driven policy π by comparing its expected T period total cost with the cost given by clairvoyant optimal policy. We refer to their delta as difference regret and formulate it as

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(C_t(y_t) - C_t(y^*))\right].$$

This problem has exactly the same setup as Shi et al. (2016). They apply a stochastic gradient algorithm to achieve $O\left(\sqrt{T}\right)$ regret rate. i.e.$\mathcal{R}_T = O\left(\sqrt{T}\right)$. It is a provably optimal regret rate with respect to scale of T. However, in this *duct* inventory problem, the number of product types n also affects regret rate significantly. A closer examination of the algorithm in Shi et al. (2016) shows that their algorithm indeed has regret rate $O\left(\sqrt{JT}\right)$, in which the $\sqrt{J}$ factor is attributed to the variance of is stochastic gradient which scale linearly with respect to J. In this paper, we propose an algorithm that achieves regret rate $O\left(\log(J)\sqrt{T}\right)$, thereby performing better when J is large. Moreover, each iteration (period) also becomes more expensive when J is large. In the policy proposed by Shi et al. (2016), each iteration requires projecting some J-dimensional vector $\tilde{y}$ onto Γ with respect to Euclidean distance. Recall that $\Gamma = \left\{y \in \mathbb{R}_+^J : \sum_{j=1}^J y^j \leq M\right\}$. So the projection operation is equivalent to solving a convex program:

$$\begin{aligned} &\min \|y - \tilde{y}\|_2 \\ \text{s.t.} \quad & y \geq 0 \\ & \sum_{j=1}^{J} y^j \leq M, \end{aligned}$$

which takes super-linear computational cost, whereas our policy requires only $O(J)$ operations in each iteration.

4.2 Mirror Descent Inventory Control Policy

When the firm has no knowledge of the true underlying distribution of D_t a priori, the goal is to find a provably good adaptive data-driven inventory control policy that yields a total cost closest to the optimal strategy.

We group periods into cycles, and our proposed policy proceeds cycle by cycle. In each cycle n, $y_{[n]}$ is the inventory target. To facilitate our analysis, we introduce a slack variable $y_{[n]}^0 = M -$

$\sum_{j=1}^{J} y_{[n]}^{j}$, and define $\tilde{y}_{[n]} := \left[y_t^0, y_t^1, \ldots, y_t^J\right] \in \mathbb{R}_+^{J+1}$. Note that $\|\tilde{y}_t\|_1 = M$. In other words, $\tilde{y}_t \in M\Delta_{J+1}$, which Δ_{J+1} denotes the $(J+1)$-simplex.

We present our multi-product mirror descent algorithm (MMD) as follows.

1. We set the first cycle starting period $\tau_1 = 1$, and let the initial inventory levels $y_{[1]} = \left[\frac{M}{n+1}, \ldots, \frac{M}{n+1}\right] \in \mathbb{R}_+^n$, i.e., $\tilde{y}_{[1]} = \left[\frac{M}{n+1}, \ldots, \frac{M}{n+1}\right] \in M\Delta_{n+1}$.

2. For cycle $n = 1, ..., N$:

   (a) At the first period of the cycle, we order up to $y_{\tau_n} = y_{[n]}$ and observe sales information $\min\left(y_{[n]}, D_{\tau_n}\right)$.

   i. We compute the gradient $\tilde{g}_{[n]}$ as follows: $\tilde{g}_{[n]}^0 = 0$ and for $j = 1, \ldots, n$,

$$\tilde{g}_{[n]}^{j} := \begin{cases} h^j + c^j & \text{for } \tilde{y}_{[n]} \geq D_{\tau_n}. \\ -p^j + c^j & \text{else.} \end{cases} \tag{4.2}$$

   ii. We do a "mirror update" as follows:

$$\tilde{y}_{[n+1]} := \frac{\tilde{y}_{[n]} \exp\left(-\eta \tilde{g}_{[n]}\right)}{Z_{[n]}} \tag{4.3}$$

   where the normalizer

$$Z_{[n]} := \frac{\left\|\tilde{y}_{[n]} \exp\left(-\eta \tilde{g}_{[n]}\right)\right\|_1}{M}.$$

   (b) In the next period $\tau_n + 1$, ideally, we would like to order up to $y_{[n+1]}$ as our next ordering target, i.e.,

$$y_{\tau_{n+1}} = \max\left(x_{\tau_{n+1}}, y_{[n+1]}\right).$$

   However, due to inventory carry-over issue and inventory capacity constrain, $y_{[n+1]}$ may not be feasible: When $y_{[n+1]}^j < x_{\tau_n+1}^j$ for some product j, if we order up to $y_{[n+1]}^j$, then the ending inventory will exceed the capacity M. Thus, in the rest of the cycle, we instead order up to $\min\left(\tilde{y}_{[n]}, \tilde{y}_{[n+1]}\right)$ until after some l_n period the on-hand inventory $x_{\tau_n + l_n} \leq \tilde{y}_{[n+1]}$. We go into the next cycle which starts at period $\tau_{n+1} := \tau_n + l_n$.

In short, the MMD policy runs cycle by cycle. In the n^{th} cycle, it first orders up to $\tilde{y}_{[n]}$ and computes $\tilde{y}_{[n+1]}$. Then in the rest of the cycle it orders up to $\min\left(\tilde{y}_{[n]}, \tilde{y}_{[n+1]}\right)$. By (4.1), the $\tilde{g}_{[n]}^j$ is indeed the gradient of C_{τ_n} as in (4.2). In each cycle we update the ordering target once, and thus cycle length will effect our updating frequency with respect to periods.

Note that in the n^{th} cycle, after the first period, we order up to $\min\left(\tilde{y}_{[n]}, \tilde{y}_{[n+1]}\right)$ repeatedly until

$x_{\tau_n+l_n} \leq \tilde{y}_{[n+1]}$. Define the duration before $x^j_{\tau_n+l_n} \leq \tilde{y}^j_{[n+1]}$ as ι^j. i.e.,

$$\iota^j := \min\left\{ s : \sum_{t=1}^{s} D^j_{\tau_n+t} \geq \tilde{y}^j_{[n+1]} - x^j_{\tau_n+1} \right\}.$$

Lemma 1. *There exits positve numbers (ν, b) such that for each j, ι^j is sub-exponential with parameter (ν, b). That is*

$$\mathbb{E}\left[e^{\lambda\left(\iota^j - \mathbb{E}\left[\iota^j\right]\right)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \textit{ for all } |\lambda| \leq \frac{1}{b}.$$

Proof. For any fixed j, consider the worst case that $\iota^j := \min\left\{ s : \sum_{t=1}^{s} D^j_t \geq M \right\}$. Note that $\mathbb{E}\left[D^j_t\right] \geq \alpha$ and $D^j_t \leq \beta$. We have

$$\begin{aligned}\alpha \leq \mathbb{E}\left[D^j_t\right] &= \mathbb{P}\left[D^j_t \leq \gamma\right]\mathbb{E}\left[D^j_t | D^j_t \leq \gamma\right] + \mathbb{P}\left[D^j_t > \gamma\right]\mathbb{E}\left[D^j_t | D^j_t > \gamma\right] \\ &\leq \mathbb{P}\left[D^j_t \leq \gamma\right]\gamma + \mathbb{P}\left[D^j_t > \gamma\right]\beta.\end{aligned}$$

If we choose $\gamma = \frac{\alpha}{2}$, we have shown that $\mathbb{P}\left[D^j_t \geq \frac{\alpha}{2}\right] \geq \frac{\alpha}{2\beta - \alpha}$. Thus,

$$\mathbb{P}[\iota > s] = \mathbb{P}\left[\sum_{t=1}^{s} D^j_t \leq M\right] \leq \mathbb{P}\left[\sum_{t=1}^{s} X_t \leq M\right]$$

where $\{X_t\}_t$ are i.i.d. random variables with distribution $X_t = \begin{cases} 0 & \text{with probability } \frac{2\beta - 2\alpha}{2\beta - \alpha} \\ \alpha/2 & \text{with probability } \frac{\alpha}{2\beta - \alpha} \end{cases}$. Note $\frac{\sum_{t=1}^{s} X_t}{\alpha/2}$ follows Binormial distribution. By Heoffding's inequality, we have

$$\mathbb{P}\left[\sum_{t=1}^{s} X_t \leq M\right] \leq \exp\left(-2\frac{\left(s\frac{2\alpha}{2\beta-\alpha} - \frac{M}{\alpha/2}\right)^2}{s}\right) \leq \exp\left(-\frac{-8\alpha}{2\beta - \alpha}\right)\exp\left(-\frac{8\alpha^2}{(2\beta - \alpha)^2}s\right).$$

Thus, ι is sub-exponential i.e. there exist (ν, b) such that

$$\mathbb{E}\left[e^{\lambda\left(\iota^j - \mathbb{E}\left[\iota^j\right]\right)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \text{ for all } |\lambda| \leq \frac{1}{b}.$$

□

From the proof, we see that (ν, b) only depend on α,β, and M. In the following analysis, we treat them as constants. Upon this lemma, the next proposition shows that the expected cycle length is of order $\log J$.

Proposition 4.1. $\mathbb{E}[l_n] \leq b\left(\log J + \frac{M+\beta}{\alpha}\right) + \frac{\nu^2}{2b^2}$ *i.e., the expected cycle length is $O(\log J)$.*

Proof. Note that $l_n = \max_{j\in[J]}\left(\iota^j\right)$. By Wald's identity, $\mathbb{E}\left[\sum_{t=1}^{\iota^i} D_t\right] = \mathbb{E}\left[\iota^i\right]\mathbb{E}\left[D_t\right]$. Note that $\sum_{t=1}^{\iota^i} D_t \le M+\beta$, and $\mathbb{E}\left[D_t\right] \ge \alpha$. We have

$$\mathbb{E}\left[\iota^i\right] \le \frac{M+\beta}{\alpha}.$$

To bound l_n, choose $\lambda = 1/b$, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in[n]}\iota^i\right] &= b\mathbb{E}\left[\log e^{\frac{1}{b}\max_{i\in[n]}\iota^i}\right] \le b\log\mathbb{E}\left[e^{\frac{1}{b}\max_{i\in[n]}\iota^i}\right] \\
&\le b\log\sum_{i\in[n]}\mathbb{E}\left[e^{\frac{1}{b}\iota^i}\right] \le b\log\sum_{i\in[n]} e^{\frac{v^2}{2b^2}+\mathbb{E}\left[\iota^i\right]} \\
&\le b\log\sum_{i\in[n]} e^{\left(\frac{v^2}{2b^2}+\frac{M+\beta}{\alpha}\right)} = b\left(\log n + \frac{M+\beta}{\alpha}\right) + \frac{v^2}{2b^2}
\end{aligned}$$

□

To prove $\mathcal{R}_T = O\left((\log J)^{1.5}\,T^{0.5}\right)$, we introduce a bridging policy $\tilde{\pi}$ which applies $y_t = \tilde{y}_{[n]}$ for all periods t in the n^{th} cycle. Note that due to the inventory carry-over, $\tilde{\pi}$ is not a feasible policy, but used for facilitating the analysis. With $\tilde{\pi}$, we decompose the regret as below.

$$\begin{aligned}
\mathcal{R}_N &= \mathbb{E}\left[\sum_{t=1}^{T} C_t(y_t) - C_t(y^*)\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N}\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y^*)\right)\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N}\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y_{[n]}) + C_{\tau_n+s}(y_{[n]}) - C_{\tau_n+s}(y^*)\right)\right] \\
&= \sum_{n=1}^{N}\mathbb{E}\left[\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y_{[n]})\right)\right] + \sum_{n=1}^{N}\mathbb{E}\left[\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{[n]}) - C_{\tau_n+s}(y^*)\right)\right]
\end{aligned}$$

To bound the second term, we rely on a Mirror descent argument.

Proposition 4.2. *Choose* $\eta = \frac{\sqrt{2\log(J+1)}}{(p+c)\sqrt{N}}$, *we have* $\sum_{n=1}^{N}\mathbb{E}\left[\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{[n]}) - C_{\tau_n+s}(y^*)\right)\right] = O\left((\log J)^{1.5}\,T^{0.5}\right)$.

Proof. By Wald's equality and Proposition 4.1, we have

$$\begin{aligned}
\sum_{n=1}^{N}\mathbb{E}\left[\sum_{s=0}^{l_n-1}\left(C_{\tau_n+s}(y_{[n]}) - C_{\tau_n+s}(y^*)\right)\right] &= \sum_{n=1}^{N}\mathbb{E}\left[l_n\right]\mathbb{E}\left[C_{\tau_n}(y_{[n]}) - C_{\tau_n}(y^*)\right] \\
&\le O(\log J)\sum_{n=1}^{N}\mathbb{E}\left[C_{\tau_n}(y_{[n]}) - C_{\tau_n}(y^*)\right]
\end{aligned}$$

Consider $C_t(y) = c \cdot y + h \cdot (y–D_t)^+ + p \cdot (y - D_t)^-$ is $(p+c)$-Lipschitz convex function on $M\Delta_{n+1}$ with respect to $\|\cdot\|_\infty$. For $\phi(y) := \sum_{i=0}^{n} y_i \log(y_i)$, $\tilde{y}_{[1]} = \left[\frac{M}{n+1}, \dots, \frac{M}{n+1}\right] = \underset{y\in M\Delta_{n+1}}{\arg\min}\phi(y)$ and we have $R^2 := \sup_{y,y'\in M\cdot_{n+1}} \phi(y) - \phi(y') \le M\log(n+1)$. We can verify (see Hazan et al. (2016)) ϕ is M-strong convex with respect to $\|\cdot\|_1$ in Δ_{n+1}. Therefore, if we take $\eta = \frac{\sqrt{2\log(J+1)}}{(p+c)\sqrt{N}}$

$$\sum_{n=1}^{N} \mathbb{E}\left[C_{\tau_n}(y_{[n]}) - C_{\tau_n}(y^*)\right] \le (p+c)\sqrt{2\log(J+1)N}. \tag{4.4}$$

To control the first term, we rely on the Lipschitz condition of C_t. □

Proposition 4.3. *Choose* $\eta = \frac{\sqrt{2\log(J+1)}}{(p+c)\sqrt{N}}$, *we have* $\sum_{n=1}^{N} \mathbb{E}\left[\sum_{s=0}^{l_n-1} \left(C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y_{[n]})\right)\right] = O\left((\log J)^{1.5} T^{0.5}\right)$.

Proof. Consider

$$\begin{aligned} C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y_{[n]}) &\le (c+p)\mathbf{1}\cdot(y_{[n]} - y_{\tau_n+s}) \\ &\le (c+p)\mathbf{1}\cdot\left(y_{[n]} - y_{[n]}e^{-2\eta(p+c)}\right) \\ &\le (c+p)\mathbf{1}\cdot 2\eta(p+c)y_{[n]} \\ &= 2\eta(p+c)^2 M. \end{aligned}$$

The second inequality is because the normalizer $Z_{[n]} \le e^{\eta(p+c)}$ and $y^j_{[n+1]} \ge y^j_{[n]}e^{-\eta(p+c)}$ for all $j \in [J]$ and the third inequality is due to $e^{-x} \ge 1 - x$.

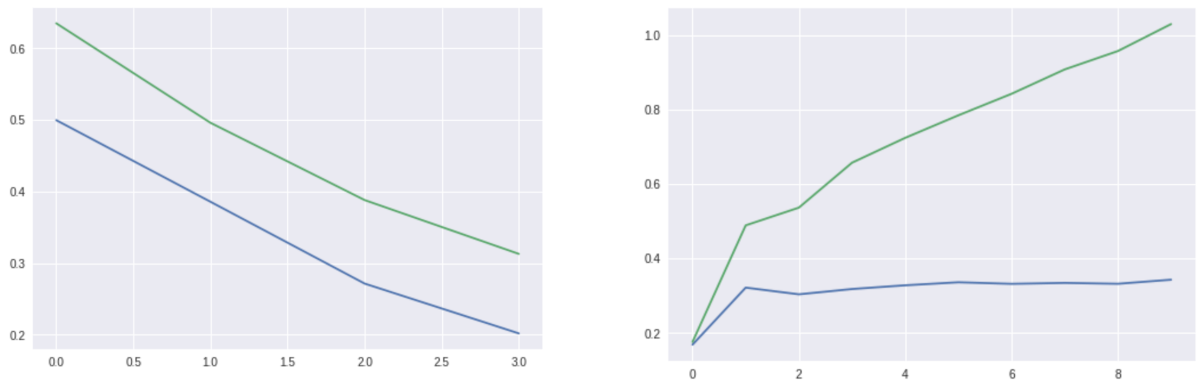
Thus, the first term

$$\begin{aligned} \sum_{n=1}^{N} \mathbb{E}\left[\sum_{s=0}^{l_n-1} \left(C_{\tau_n+s}(y_{\tau_n+s}) - C_{\tau_n+s}(y_{[n]})\right)\right] &\le N2\eta(p+c)^2 M\mathbb{E}[l_n] \\ &\le N2\eta(p+c)^2 MO(\log J) \\ &= O\left((\log J)^{1.5} N^{0.5}\right) \end{aligned}$$

□

4.3 Simulation

In the left plot, we fix a small N, and we see that our mirror descent based algorithm has slightly better performance than the stochastic gradient based algorithm. In the right plot, we fix T, let N grow, and we see that our policy has much better performance.



(a) Fix N

(b) Fix T

Figure 4.1: SGD(Green) v.s. Mirror Descent(Blue)

4.4 Appendix

Theorem 4.2. *Let $\{f_t\}_{t=1}^T$ be a sequence of G-Lipschitz convex function on some norm space $\mathcal{D}$ with norm $\|\cdot\|$. Let ϕ be a ρ-strong convex function on $\mathcal{D}$ with respect duel norm $\|\cdot\|^*$ with diameter square $R^2 := \sup_{x,x'\in\mathcal{D}} \phi(x) - \phi(x')$. If $x_1 = \underset{x\in\mathcal{D}}{\arg\min}\phi(x)$ and for $t = 1,\ldots,T-1$, x_{t+1} follows the mirror descent update*

$$x_{t+1} = \underset{x\in\mathcal{D}}{\arg\min} D_\phi\left(x; \nabla^{-1}\phi(\nabla\phi(x_t) - \eta\nabla f_t(x_t))\right),$$

where D_ϕ is the Bergman divergence. Then

$$\sum_{t=1}^{T}[f_t(x_t) - f_t(x)] \le \frac{R^2}{\eta} + \frac{\eta}{2\rho}G^2T$$

for any $x \in \mathcal{D}$.

For detailed proof, see Lee and Vempala (2019).

CHAPTER 5

Marrying Stochastic Gradient Descent with Bandits: Learning Algorithms for Inventory Systems with Fixed Costs

5.1 Introduction

The periodic-review stochastic inventory control problem with fixed cost is perhaps the most fundamental problem in the theory and practice of inventory management (cf. Zipkin (2000) and Simchi-Levi et al. (2014)). A firm needs to make sequential inventory replenishment decisions over the planning horizon under stochastic demand, with the objective to minimize the total ordering, holding, and lost-sales penalty costs. The ordering cost in each period typically consists of two components, namely, the variable cost and the fixed cost. The variable cost is the ordering quantity times the per-unit ordering cost, whereas the fixed cost is a constant $K > 0$ whenever a positive order is placed. Undoubtedly, fixed costs arise in many real-life scenarios, and reflect the fact that ordering, production, and transportation in large quantities lead to economies of scales.

It is well-known in the literature that the problems with fixed cost are much harder to analyze than those without, because the newsvendor cost (including the holding and lost-sales penalty costs) is convex in the decision space while the ordering cost is effectively concave in the presence of fixed cost. The celebrated paper by Scarf (1960) proved that a so-called (s, S) policy is optimal for such a problem over a finite horizon via the elegant notion of K-convexity. Subsequently, Iglehart (1963) and Zheng (1991) established the optimality of the (s, S) policy for the infinite horizon counterpart problem. The (s, S) policy admits a very simple structure, i.e., the firm should place an order to bring the inventory position back to the order-up-to level S whenever the on-hand inventory drops below the triggering level s. Efficient searching heuristics for the optimal policy have also been proposed by Veinott Jr and Wagner (1965) and Zheng and Federgruen (1991).

To this date, almost all the papers on this fundamental topic assume that the stochastic demand processes are given as input to the models, and the inventory replenishment decisions are made with full knowledge of the demand distribution. However, in practice, the underlying demand distribution may not be available to the firm *a priori*. The firm may collect past sales over time to estimate the demand distribution. However, since the realized sales in a period are the minimum of the actual demand and the on-hand inventory level, this demand information is *censored* (cf. Huh and Rusmevichientong (2009)). This raises a natural and important research question as to how to devise an efficient and effective learning algorithm that only uses the sales data collected over time to minimize the cumulative expected newsvendor cost.

5.1.1 Main Results and Contributions

We propose the first nonparametric learning algorithm, termed the (δ, S) policy, for the periodic-review stochastic inventory system with fixed cost under censored demand, where $\delta := S - s$ is defined as the inventory gap between the order-up-to level S and the triggering level s. The performance is measured using the notion of *cumulative regret* (or simply *regret*), which is the difference between the expected cost of our proposed policy and that of the clairvoyant optimal policy over $T \geq 1$ periods. We show that under mild assumptions, the regret of the (δ, S) policy is $O(\log T \sqrt{T})$, which is provably optimal up to some logarithmic factor.

Our result significantly contributes to the growing nonparametric inventory learning literature, first initiated by the renowned result by Huh and Rusmevichientong (2009) for the classical inventory systems and followed up by many other results surveyed in §5.1.2. However, there are very few results thus far on the simplest setting with fixed cost, despite its clear importance and practical relevance. Perhaps the main reason that hinders such a progress is that the objective function is not jointly convex in S and s (or S and δ) with the fixed cost $K > 0$, rendering the direct adaptation of the online *Stochastic Gradient Descent* (SGD) method (Shalev-Shwartz et al. (2012)) prohibitive.

We shall summarize our high-level approaches and major contributions below.

1. **Merging the first-order and zeroth-order optimizations.** Although the objective function is not jointly convex in S and δ, for any *fixed* inventory gap δ, the objective function (in terms of the so-called "cycle cost" defined as the cost between successive attaining of the target level S) is convex in S. This motivates us to design a (δ, S) learning algorithm that searches for the optimal policy by integrating a first-order optimization method (running a sub-exponential SGD on S) with a zeroth-order optimization method (running bandit controls on δ).

   The high-level idea is as follows. We discretize the space of inventory gap δ, and maintain a what-we-call *active set* that adaptively keeps track of all favorable candidate policies. At each iteration, for each fixed inventory gap δ within the active set, we perform a SGD step on the order-up-to level S and complete a cycle (i.e., waiting for all active policies to hit S

again). We then utilize the collected censored demands to simulate the performance of all other candidate policies within the active set. Finally, we prune the active set by eliminating the "unfavorable" candidate policies by examining their hitherto empirical performance and carefully controlling the confidence bound. Our aim is to retain the optimal policy with an overwhelmingly high probability, while gradually shrinking the active set. The regret analysis needs to bound the SGD sub-optimality loss and the pruning loss *simultaneously*.

We note that the previous approaches used in Huh et al. (2009); Huh and Rusmevichientong (2009); Zhang et al. (2018, 2019) rely heavily on the convexity of the objectives, and therefore *need not* maintain and adaptively prune an active set of all favorable policies. Taking into account of fixed cost, just as every other paper in inventory theory (be it learning or not), changes the landscape of the problem and spurs a new methodological development.

2. **Simulation of all active policies (SAAP).** To achieve a tight regret bound and speed up the computation, one can hope to use the censored demand information collected by running one particular (δ, S) policy to simulate the performance of other policies. The difficulty lies in the fact that, due to demand censoring, a randomly chosen policy may not grant sufficient demand information to simulate all the feasible policies within the active set. By judiciously choosing the "information maximizing" policy at the beginning of each cycle and adhering to this policy sufficiently long, we ensure that the collected demand information is sufficient for simulating all active policies. This idea is central to our algorithmic design and regret analysis because of the special structure of the problem.

3. **Bounding the regret loss through multiple bridging policies.** Bounding the cumulative regret of the (δ, S) learning algorithm would necessitate the following bridging policies.

   (a) *Optimal policy on the grid.* We equally discretize the domain of the inventory gap δ to initialize our bandit-control-based algorithm. By proving the smoothness of the transformed objective function, we are able to control the discretization loss.

   (b) *Optimal policy on* $\{\delta^n\}$. Our algorithm suggests a sequence of δ^n decisions based on their hitherto empirical performance. We show that $\{\delta^n\}$ converges to the optimal δ^* and bound the loss between the optimal policy on $\{\delta^n\}$ and the optimal policy on the grid.

   (c) *SGD policy.* For each fixed δ, its corresponding target level S is calculated via SGD. We prove that the sequence of S converges to the δ-specific optimal S^* uniformly across δ.

   (d) *Implemented learning policy.* Due to positive inventory carryover, the above SGD policy may not be implementable, e.g., when the updated target level S is below the current inventory position. We develop a queueing system argument to bound the loss due to this discrepancy between the suggested SGD policy and the implemented learning policy.

4. **Technical results of independent interest.** As a by-product of our regret analysis, we develop two useful technical results that are of independent interest. First, we show the Lipschitz continuity of the hitting time of an ascending random walk. Second, we develop a new high probability bound for sub-exponential SGD that could be applied in more general settings.

5. **Connection with the general Lipschitz bandits.** By treating cycle costs as period costs and ignoring the inventory carryover constraints, our (δ, S) policy could be regarded as a two-dimensional continuum-armed bandit problem in which Bubeck et al. (2011) and Kleinberg et al. (2008) proposed policies that achieve the optimal regret $O(T^{\frac{d+1}{d+2}})$ where d is the dimension of the decision variable space by only assuming Lipschitz continuity on the objective function. Our (δ, S) control is also two-dimensional, and hence the bulk of this work is to utilize the partial convexity structure to lower the regret from $O(T^{3/4})$ to $O(\log T\sqrt{T})$. To achieve this, we need to analyze the empirical estimator of the objective value, develop a new high probability bound for sub-exponential SGD, and also carefully choose the confidence size to prune our active set in each iteration of the learning algorithm.

In summary, our (δ, S) learning algorithm and performance analysis involve several innovative ideas that ultimately enable the integration of SGD and bandit controls, achieving the best of both worlds. Going forward, we believe that our framework opens up many doors for multi-dimensional decision making in which the objective function has only some partial convexity or concavity structures, as it draws the strength of both SGD and bandit controls in a non-trivial fashion.

5.1.2 Literature Review

Our work is closely related to two streams of literature discussed below.

Inventory systems with fixed costs. Given the complete information on demand distribution, there has been a large body of literature on inventory systems with fixed costs. The seminal papers Scarf (1960) and Veinott Jr (1966) characterized the optimal ordering decision as a state-dependent (s, S) policy. Iglehart (1963) and Zheng (1991) established the optimality of the stationary (s, S) policy for the infinite horizon counterpart problem. Sethi and Cheng (1997) and Gallego and Özer (2001) showed that the structural result continues to hold under the model with Markov-modulated demand and the model with advance demand information, respectively.

Besides the structure of optimal policies, substantial efforts have also been spent on designing efficient heuristics. With i.i.d. demands, Federgruen and Zipkin (1984) presented an iterative algorithm that converges to the optimal stationary policy over an infinite horizon. Zheng and Federgruen (1991) computed the optimal policy for a continuous-reviewed system with a renewal demand process. Gallego and Özer (2001) and Özer and Wei (2004) developed efficient algorithms for models with advance demand information. Gavirneni (2001) proposed a heuristic for finding the supplier's non-stationary capacitated inventory control policy given the retailer had adopted an (s,S) policy. Guan and Miller (2008) and Huang and KüçüKyavuz (2008) gave polynomial-time exact algorithms for the uncapacitated lot-sizing problem given that the stochastic programming scenario tree is polynomially representable. Levi and Shi (2013) and Shi et al. (2014) gave approximation algorithms with worst-case performance guarantees under general demand models.

Beyond the basic model with fixed costs, there have been streams of research focusing on incorporating other factors (together with fixed cost), including, but not limited to, capacitated problems (e.g., Chen (2004); Chen and Lambrecht (1996); Gallego and Scheller-Wolf (2000)), joint pricing and inventory control problems (e.g., Chen and Simchi-Levi (2004a,b); Chen et al. (2006); Feng (2010); Hu et al. (2018); Huh and Janakiraman (2008); Pang et al. (2012)), quantity dependent fixed costs (e.g., Caliskan-Demirag et al. (2012); Chao and Zipkin (2008), and joint replenishment problems (e.g., Cheung et al. (2016); Khouja and Goyal (2008); Nagarajan and Shi (2016)).

Learning algorithms for inventory systems. We can roughly divide learning algorithms into two categories, namely, *parametric* algorithms and *nonparametric* algorithms, depending on the firm's information structure. In the former category, the firm forms a prior belief about the demand distribution, and repeatedly update the parameters of the distribution with new demand information. This type of Bayesian approach was first adopted by Iglehart (1964); Murray and Silver (1966); Scarf (1959) and Azoury (1985). With demand censoring where the firm can only observe realized sales, the easier case to handle is the perishable inventory (with product lifetime being one) where the excess inventory in the current period does not carry over to the next. Lu et al. (2005, 2008) established that the optimal stocking quantity is higher than the myopic solution. The intuition is that by stocking higher, it is more likely that we can obtain more accurate, uncensored demand information, which is useful for future decisions. Unfortunately, this upper bound result does not hold for nonperishable inventory in general (see Chen and Plambeck (2008)). In a separate vein, Liyanage and Shanthikumar (2005) and Chu et al. (2008) developed an approach called operational statistics to find a decision rule that maximizes the performance uniformly for all possible values of the unknown demand parameters.

In contrast to the parametric approaches, this paper belongs to the growing body of nonparametric learning literature. Burnetas and Smith (2000) developed a learning algorithm for the repeated newsvendor problem with pricing. With demand censoring, Huh and Rusmevichientong (2009) proposed a gradient descent based algorithm for the classical multiperiod inventory system with censored demand. Besbes and Muharremoglu (2013) examined the discrete demand case and showed that active exploration is needed. Huh et al. (2011) proposed another adaptive algorithm based on Kaplan-Meier estimator. Subsequently, there has been an active stream of research devising learning algorithms for various models, namely, capacitated inventory systems (Chen et al. (2018d); Shi et al. (2016)), perishable inventory systems (Zhang et al. (2018)), lost-sales inventory systems with lead times (Huh et al. (2009); Zhang et al. (2019)), and joint pricing and inventory control (Chen et al. (2018a,b)). Another popular nonparametric approach in the inventory literature is sample average approximation (SAA) (e.g., Kleywegt et al. (2002); Levi et al. (2015, 2007)) 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.

All the learning algorithms surveyed above did not model fixed cost, mainly due to the loss of convexity structure. Existing methods cannot be readily employed or adapted to this setting.

5.1.3 Paper Organization and General Notation

The remainder of this paper is organized as follows. We formulate our problem in §5.2. We describe the (δ, S) learning algorithm in §5.3. We carry out the regret analysis in §5.4. We show some computational performance in §5.5. Finally, we conclude and point out several future directions in §5.6.

For any $x \in \mathbb{R}$, $x^+ = \max\{x, 0\}$, $x^- = \max\{-x, 0\}$. The indicator function $\mathbb{1}(A)$ takes value 1 if A is true and 0 otherwise. The notation ":=" stands for "defined as". For any integer $C \geq 1$, the vector $[C] = \{1, 2, \ldots C\}$. The projection operator $\mathbf{Proj}_{[a,b]}(x) = \max\{a, \min\{x, b\}\}$. We use LHS and RHS as abbreviations for the "left-hand side" and the "right-hand side" of an equation, respectively, and pdf and cdf as abbreviations for the "probability distribution function" and the "cumulative distribution function", respectively.

5.2 The Periodic-Review Stochastic Inventory System with Fixed Cost

We formally describe the periodic-review stochastic inventory system with fixed costs under *lost sales* and *censored demand.* Let $t \in \{1, 2, \ldots\}$ represent the time period, which is indexed forward. We denote the demand in period t by D_t, and assume that D_t, $t = 1, \ldots, T$, are independent and identically distributed (i.i.d.) continuous random variables across periods. In each period $t = 1, \ldots, T$, four types of incurred costs include (1) a per-unit ordering cost c for ordering product at the beginning of period t, (2) a *fixed ordering* cost K that is incurred when the ordering quantity in period t is positive, (3) a per-unit holding cost h for holding excess inventory from period t to $t+1$, (4) a per-unit lost-sales penalty p that is incurred when the demand at the end of period t is unsatisfied. Unsatisfied demand units are *lost* and *unobserved* due to demand censoring. The order lead time is assumed to be zero. We remark here that even for the "non-learning" problem (where the exact demand distribution is available), adding a positive order lead time to a lost-sales model with no fixed cost makes the characterization of optimal policies intractable (see Zipkin (2008)), let alone with fixed costs.

5.2.1 System Dynamics

In each period of time t, the sequence of events is as follows:

1. At the beginning of each period t, the firm observes the beginning on-hand inventory level x_t.

2. The firm makes an ordering decision $q_t \geq 0$. The ending on-hand inventory level (after receiving the order q_t) becomes $y_t = x_t + q_t$.

3. Then the demand D_t is realized to be d_t.

4. If the firm places a positive order quantity, i.e., $q_t > 0$, it incurs a fixed ordering cost K and a variable ordering cost cq_t. All the outstanding inventories incur a per-unit holding cost h, and all the unsatisfied demand units are *lost* with a per-unit lost-sales penalty cost p. Note that $p > c$ in the lost-sales model (see Zipkin (2000)). Hence, the total cost for period t is given by

$$C_t(x_t, q_t, d_t) = K\mathbb{1}(q_t > 0) + cq_t + h(x_t + q_t - d_t)^+ + p(x_t + q_t - d_t)^-. \tag{5.1}$$

5. Then, the inventory carried over to the next period $t+1$ is given by $x_{t+1} = [x_t + q_t - d_t]^+$.

Note that all x_t, y_t, q_t, C_t are policy-dependent and should be written as $x_t^\pi, y_t^\pi, q_t^\pi, C_t^\pi$ for a feasible policy π. For brevity, we will make the dependency on π implicit whenever there is no ambiguity.

5.2.2 Objective and Assumptions

Let $\mathcal{F}_t$ denote the set of all historical information collected up to the beginning of period t, which includes the past censored demand observations and the past ordering decisions.

Definition 3. *A feasible* learning *policy π is a sequence of functions $\{\pi_t\}_{t=1}^{\infty}$ such that $y_t = \pi_t(x_t, \mathcal{F}_t)$ mapping the state, the beginning inventory x_t and $\mathcal{F}_t$, to a decision, the ending inventory y_t that satisfies $y_t \geq x_t$ (i.e., the ordering decision $q_t = y_t - x_t \geq 0$).*

Our objective is to find a policy π that minimizes the *long-run average expected cost*

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi}\right], \tag{5.2}$$

where C_t^{π} is the period-t cost by running π. In this paper, we assume that the demand distribution is unknown to the firm *a priori*, and the firm can only make adaptive decisions based on the censored demands observed over time. Thus, the notion of *regret* (or cumulative regret) from online learning (see Shalev-Shwartz et al. (2012)) is a sound performance measure of π, which is defined to be the difference in total cost between π (that only makes use of censored demands collected over time) and a *clairvoyant* optimal policy (that knows the actual demand distribution *a priori*).

Definition 4. *For a feasible learning policy π, the T-period* regret *of π is*

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi}\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi^*}\right],$$

where π^* is the optimal policy that minimizes the long-run average expected cost (5.2).

Assumption 5. *Throughout this paper, we make the following mild assumptions.*

1. *Demands $\{D_t\}_{t=1}^{\infty}$ are i.i.d. across all time period t. We use a time-generic symbol D to denote the distribution, i.e., $D_t \overset{d}{=} D$.*

2. *The probability density function $f(\cdot)$ of demand D is bounded, i.e., $f(d) \leq \rho$ for all $d \geq 0$ for some constant $\rho > 0$.*

3. *The warehouse storage capacity is β (and hence the clairvoyant optimal target level $S^* \leq \beta$).*

We remark that all the above assumptions are very mild in the inventory learning literature (see, e.g., Huh et al. (2009); Huh and Rusmevichientong (2009); Zhang et al. (2018, 2019)).

5.2.3 Clairvoyant Optimal Policy – A (δ, S) Policy

To set the baseline for seeking a good π, we first characterize the clairvoyant optimal policy when the demand distribution D are given *a priori*. For ease of the description and analysis of our learning algorithm later, we slightly transform the conventional (s, S) policy to an equivalent (δ, S) policy where $\delta := S - s$ is what-we-term *inventory gap*, which is formalized below.

Definition 5. *We call an inventory control policy π an (s,S)* policy*, if the order-up-to level*

$$y_t = \pi_t(x_t) = \begin{cases} S & \text{if } x_t \leq s, \\ x_t & \text{if } x_t > S. \end{cases}$$

Equivalently, by setting $\delta := S - s$, we call an inventory control policy π a (δ,S) policy*, if*

$$y_t = \pi_t(x_t) = \begin{cases} S & \text{if } x_t \leq S - \delta, \\ x_t & \text{if } x_t > S. \end{cases}$$

That is, whenever the on-hand inventory x_t falls below the triggering level $S-\delta$, the firm places an order of $q_t = S - x_t$ to bring the inventory up to S. Note that $q_t \geq \delta$ whenever $q_t > 0$, and that is the reason we term δ the"inventory gap". The next result re-emphasizes that such a simple structure is optimal for the clairvoyant problem, and a very concise proof was given by Zheng (1991).

Theorem 5.1 (Zheng (1991))**.** *When the demand distribution is known a priori, there exists a pair of inventory gap $\delta^* \in \mathbb{R}$ and order-up-to level $S^* \in \mathbb{R}$ such that the (δ^*,S^*) policy is optimal, i.e., the (δ^*,S^*) policy minimizes the long-run average expected cost* (5.2).

5.3 A (δ,S) Learning Algorithm

For a fixed (δ,S) policy, we record the time periods in which the after-ordering inventory y_t hits the inventory target S, i.e., we use $\{\tau_i\}_{i=1}^{\infty}$ to denote these reordering time periods, i.e.,

$$\tau_1 = \inf\{t \geq 1 : y_t = S\}, \quad \tau_{i+1} = \inf\{t > \tau_i : y_t = S\}, \text{ for } i = 1,2,\ldots \tag{5.3}$$

We call the time interval $[\tau_i,\tau_{i+1})$ the i^{th} *cycle* of the system. Let $L_i^{(\delta,S)}$ and $H_i^{(\delta,S)}$ denote the length and the cost of the i^{th} cycle, respectively, i.e.,

$$L_i^{(\delta,S)} := \tau_{i+1} - \tau_i, \quad H_i^{(\delta,S)} := c\,(x_{\tau_i} - x_{\tau_{i+1}}) + \sum_{t=\tau_i}^{\tau_{i+1}-1} C_t. \tag{5.4}$$

Note that $H_i^{(\delta,S)}$ includes the newsvendor cost over $[\tau_i,\tau_{i+1})$ and the ordering cost at τ_{i+1}. For notational convenience, we use $L(\delta,S)$ and $H(\delta,S)$ to denote the (time-generic) length and cost of a random cycle, respectively. Note that $L(\delta,S)$ is, in fact, independent of S, and therefore we shall use $L(\delta)$ for $L(\delta,S)$ interchangeably in the remainder of this paper.

Figure 5.1 illustrates a random cycle of a fixed (δ,S) policy. It is evident that the inventory level process regenerates at each τ_i ($i = 1,2,\ldots$) when the after-ordering inventory level increases

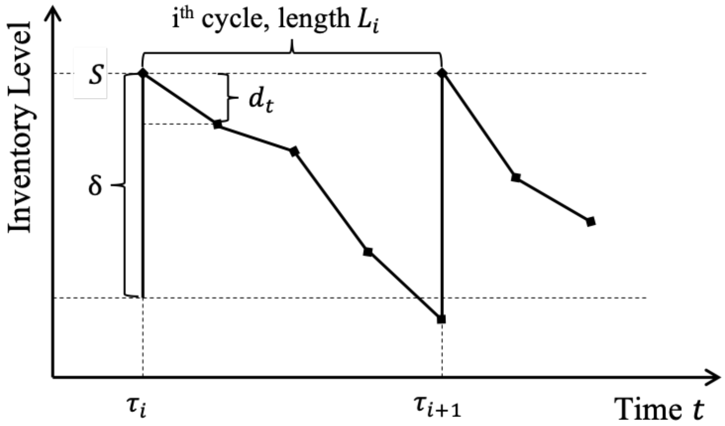


Figure 5.1: A random cycle with respect to a fixed (δ, S) policy

to S (see Zheng and Federgruen (1991)). By Assumption 5, we know that $\mathbb{E}|H(\delta, S)| < \infty$, and then by applying the Renewal Reward Theorem (see Ross (1996)), we have

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{(\delta,S)}\right] = \frac{\mathbb{E}[H(\delta, S)]}{\mathbb{E}[L(\delta, S)]}. \tag{5.5}$$

In essence, (5.5) says that the long-run average expected cost of a fixed (δ, S) policy equals the expected cycle cost divided by the expected cycle length. Hence, for the clairvoyant problem, it follows that the optimal policy π^* is given by

$$\pi^* = (\delta^*, S^*) \in \arg\min_{(\delta,S)} \frac{\mathbb{E}[H(\delta, S)]}{\mathbb{E}[L(\delta, S)]}. \tag{5.6}$$

5.3.1 Transforming the Objective

Following the discussion above, $H(\delta, S)$ and $L(\delta, S)$ are the key variables used to evaluate the performance of a fixed (δ, S) policy. However, due to the censored demand, when a lost sale occurs during a cycle, the cycle cost $H(\delta, S)$ is not fully observable. Then we seek an alternative measurement of that cost that is *observable* to the firm. Note that the censored part of the cost $C_t(x_t, q_t, d_t)$ in (5.1) is $p(x_t + q_t - d_t)^-$, which is not observable whenever a lost sale occurs. Nevertheless, we can decompose it as

$$p(x_t + q_t - d_t)^- = pd_t - p\min(x_t + q_t, d_t).$$

The above simple transformation is *crucial* for our analysis by observing that (a) the first term pd_t is independent of any feasible policy π, and (b) the second term $p\min\{x_t + q_t, d_t\}$ is policy-dependent but observable. Thus, we define what-we-call *pseudo cost* by dropping the first term

$$\tilde{C}_t(x_t, q_t, d_t) := C_t(x_t, q_t, d_t) - pd_t = K\mathbb{1}[q_t > 0] + cq_t + h(x_t + q_t - d_t)^+ - p\min(x_t + q_t, d_t). \quad (5.7)$$

For a fixed (δ, S) policy, we define the *cycle pseudo cost* to include all the newsvendor cost over the interval $[\tau_i, \tau_{i+1})$ and the ordering cost at τ_{i+1}, i.e.,

$$G_i^{(\delta,S)} := c(x_{\tau_i} - x_{\tau_{i+1}}) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \tilde{C}_t.$$

Again, for notational convenience, we use $G(\delta, S)$ to denote the (time-generic) cycle pseudo cost of a random cycle. The same renewal argument for deriving (5.5) implies that

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \tilde{C}_t^{(\delta,S)}\right] = \frac{\mathbb{E}[G(\delta,S)]}{\mathbb{E}[L(\delta,S)]}. \quad (5.8)$$

Also, we note that

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t^{(\delta,S)}\right] - \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \tilde{C}_t^{(\delta,S)}\right] = p\mathbb{E}[D].$$

The difference between the long-run average cost and pseudo cost is $p\mathbb{E}[D]$, which is independent of any chosen feasible policy, and therefore an equivalent objective to (5.6) is given by

$$\pi^* = (\delta^*, S^*) \in \arg\min_{(\delta,S)} \frac{\mathbb{E}[G(\delta,S)]}{\mathbb{E}[L(\delta,S)]}. \quad (5.9)$$

For notational convenience, we let $V(\delta, S) := \mathbb{E}[G(\delta,S)]/\mathbb{E}[L(\delta,S)]$, and the optimal value under the optimal policy is $V^* := V(\delta^*, S^*)$.

5.3.2 Properties of the Transformed Objective

We shall establish some desirable properties of $V(\delta, S)$, such as convexity and smoothness, in Theorem 5.2. For better flow of the paper, the proofs for this section are deferred to the Appendix.

Lemma 2. *Let $\gamma := K + (h + c + p)\beta$. Then the pseudo cost $|\tilde{C}_t| \leq \gamma$ for all $t \geq 1$.*

Because $G(\delta, S)$ is cycle pseudo cost and $L(\delta, S)$ is cycle length, following Lemma 2, we immediately have $|G(\delta, S)/L(\delta, S)| \leq \gamma$. To establish the smoothness of $V(\delta, S)$, we prove a non-trivial technical result, which asserts that the hitting time of an ascending random walk is Lipschitz in the hitting target level. The proof is of independent interest.

Lemma 3 (Lipschitz Hitting Time of Ascending Random Walk)**.** *Suppose $\{D_t\}_{t=1}^{\infty}$ is a sequence of i.i.d. positive random variables with densities bounded by constant ρ. Define an ascending random walk $W_0 = 0$, $W_t = \sum_{i=1}^{t} D_i$. For any $\delta > 0$, let $L(\delta)$ be the hitting time to the interval $[\delta, \infty)$, i.e., $L(\delta) = \min\{t : W_t \geq \delta\}$. Then, $\mathbb{E}[L(\delta)]$ is Lipschitz in δ with Lipschitz constant 6ρ.*

With the aid of the above two lemmas, we shall show that $V(\delta, S)$ has *convexity* and *smoothness* properties that are essential for building our (δ, S) learning algorithm.

Theorem 5.2 (Stucture of the Objective Function $V(\delta,S)$)**.** *When $0 \leq \delta \leq S \leq \beta$, our objective function $V(\delta,S)$ has the following properties.*

1. *For any fixed δ, $V(\delta,S)$ is Lipschitz and convex in S. Moreover, the Lipschitz constant can be chosen independent of δ.*
2. *For any fixed δ, given a cycle with length L and an ending inventory level x_{L+1}, we have*

$$\tilde{\nabla} := \begin{cases} hL, & \text{if } x_{L+1} > 0 \\ h(L-1) - p + c, & \text{if } x_{L+1} = 0 \end{cases} \tag{5.10}$$

is an unbiased stochastic (S-partial) gradient for $\mathbb{E}[G(\delta,S)]$, i.e., $\nabla_S \mathbb{E}[G(\delta,S)] = \mathbb{E}\left[\tilde{\nabla}\right]$.

3. *Define the objective function in δ with the optimized S for each δ by*

$$V^*(\delta) := \min_{S \in [\delta,\beta]} V(\delta,S). \tag{5.11}$$

Then $V^*(\delta)$ is Lipschitz in δ.

5.3.3 Description of the (δ, S) Learning Algorithm

The proposed learning (δ,S) algorithm is the first nonparametric learning algorithm for stochastic inventory system with fixed costs. This algorithm integrates the zeroth- and first-order optimization and simulation techniques. Leveraging on this innovative framework, our algorithm can achieve provably optimal regret rate up to a logarithmic factor (see the performance analysis in §5.4).

Parameters. Let J be the number of discrete inventory gaps, η_n be the stochastic gradient descent step size of the n^{th} iteration, Δ^n be the so-called *confidence size* of the n^{th} iteration. We will specify how to choose the optimal values of these parameters later in our main analysis.

Initialization.

(0a) We discretize the feasible region of inventory gap δ, namely, $[0,\beta]$, into an equal spacing set of J number of inventory gaps $\{\delta_1, \ldots, \delta_J\}$ where $\delta_j = j\beta/J$.

(0b) For inventory gap δ_j, we initialize the target inventory level $S_j^1 \in [\delta_j, \beta]$ arbitrarily.

(0c) The algorithm maintains a what-we-call *active set* $\mathcal{A}^n$ that contains the favorable candidates of optimal solution after the n^{th} iteration. We initialize $\mathcal{A}^1 = \{1, \ldots, J\}$. As the learning algorithm proceeds and n increases, the set $\mathcal{A}^n$ decreases. (We are removing the unlikely candidates gradually when information is sufficient.)

(0d) For each $j \in \mathcal{A}^n$, we also keep track of the cumulative cycle pseudo cost $\hat{G}_j^{n-1}$ and the cumulative cycle length $\hat{L}_j^{n-1}$ over the last $n-1$ iterations. We initialize $\hat{G}_j^0 = 0$ and $\hat{L}_j^0 = 0$ for $j = 1, 2, \ldots, J$.

Main Loop. The algorithm proceeds in *epochs* where each epoch consists of one or more complete cycles. We index the epochs by $n = 1, 2, \ldots,$ and each epoch corresponds to an exact iteration of the algorithm below. The main idea of the algorithm proceeding in epochs is to ensure that the collected (censored) demand information within each epoch n is sufficient for *simulating* all the candidate policies in the active set $\mathcal{A}^n$ *in parallel*.

At the beginning of each epoch n, we observe the on-hand inventory, denoted by x^n. We perform the following *crucial* operation termed *Simulation of All Active Policies* (SAAP).

(1a) Find the maximum target level S_j^n among policies $j \in \mathcal{A}^n$ and record its index

$$j^n := \arg\max_{j \in \mathcal{A}^n} S_j^n. \tag{5.12}$$

In other words, the policy $(\delta_{j^n}, S_{j^n}^n)$ has the highest target order-up-to level among all "favorable candidates" within the active set $\mathcal{A}^n$.

(1b) Compare S_{j^n} with the epoch beginning inventory level x^n to decide whether S_{j^n} is feasible. If $S_{j^n} > x_n$, we choose to implement the policy (δ_{j^n}, S_{j^n}); otherwise, we choose to implement the policy (δ_{j^n}, x^n). In other words, for the n^{th} epoch, we choose to implement the policy

$$(\delta^n, S^n) := (\delta_{j^n}, \max(x^n, S_{j^n}^n)). \tag{5.13}$$

(1c) Find the maximum inventory gap, denoted by $\bar{\delta}^n$, among the active set $\mathcal{A}^n$. That is,

$$\bar{\delta}^n := \max_{j \in \mathcal{A}^n} \delta_j. \tag{5.14}$$

We run the policy (δ^n, S^n) in (5.13) for some complete cycles until the *cumulative* demand for the n^{th} epoch exceeds the maximum inventory gap $\bar{\delta}^n$. We then go to the next epoch.

(1d) Leverage the (censored) demand collected within the n^{th} epoch to *simulate* all the candidate policies $j \in \mathcal{A}^n$ in parallel, and collect their respective simulated cycle pseudo cost G_j^n and cycle length L_j^n and stochastic gradient $\tilde{\nabla}_j^n$ defined in (5.10). We will discuss how and why such a simulation operation works in the next subsection.

With the simulated information gathered above by SAAP, we carry out the following updates.

(2a) Update S_j^n for each $j \in \mathcal{A}^n$ by a stochastic gradient step (the first-order optimization):

$$S_j^{n+1} = \mathbf{Proj}_{[\delta_j, \beta]}\left(S_j^n - \eta_n \tilde{\nabla}_j^n\right),$$

where the stochastic gradient $\tilde{\nabla}_j^n$ defined in (5.10).

(2b) Update the cumulative cycle pseudo cost and cycle length for each $j \in \mathcal{A}^n$:

$$\hat{G}_j^n = \hat{G}_j^{n-1} + G_j^n, \qquad \hat{L}_j^n = \hat{L}_j^{n-1} + L_j^n. \tag{5.15}$$

(2c) Update the active set $\mathcal{A}^n$ by a bandit-like step (the zeroth-order optimization):

$$\mathcal{A}^{n+1} = \left\{ j \in \mathcal{A}^n : \frac{\hat{G}_j^n}{\hat{L}_j^n} - \min_{j'} \frac{\hat{G}_{j'}^n}{\hat{L}_{j'}^n} \leq \Delta^n \right\}. \tag{5.16}$$

The main idea is to *prune and refine* the active set $\mathcal{A}^n$ based on their empirical performances gauged by a confidence metric Δ^n. That is, with very high probability, suboptimal policies will be gradually removed from this active set. We go to the $(n+1)^{th}$ epoch, and repeat.

This concludes the description of our (δ, S) learning algorithm. For the convenience of practical implementation, we also provide a detailed pseudo code in Algorithm 5.1.

5.3.4 Main Ideas of SAAP

A pivotal step in the algorithm is the *Simulation of All Active Policies* (SAAP). We shall discuss why the above specified rules can indeed collect sufficient demand information from applying one particular "information maximizing" policy to simulate all the rest within an active set.

Consider two policies (δ_1, S_1) and (δ_2, S_2) with $\delta_1 \geq \delta_2$ and $S_1 \geq S_2$. It is clear that the demands collected from a cycle of the policy (δ_1, S_1) can be used to simulate a cycle of policy (δ_2, S_2). An example is shown in Figure 5.2. We simply re-use and "shift down" the sample path obtained from running (δ_1, S_1) to simulate (δ_2, S_2) and apply early stopping when the cumulative demand exceeds δ_2, which is always feasible. Thus, following the above logic, in the n^{th} epoch if we could

Algorithm 5.1 A (δ, S) Learning Algorithm

```
Let $J$ = # discrete $\delta$'s, and $\eta_n$ = step size, and $\Delta^n$ = confidence size.    ▷ Parameters
Initialize the active set $\mathcal{A}^1 = \{1, \dots, J\}$;    ▷ Initialization
For $j \in \mathcal{A}^1$, define $\delta_j = \frac{j}{J}\beta$ and assign $S_j^1 \in [\delta_j, \beta]$ arbitrarily; set the cumulative cycle pseudo cost $\hat{G}_j^0 = 0$ and the cumulative cycle length $\hat{L}_j^0 = 0$.
for $n = 1, 2, \dots$ do    ▷ Main Loop
    Let $x^n$ be the epoch beginning inventory.
    Let the index $j^n = \arg\max_{j \in \mathcal{A}^n} S_j^n$ (with the maximum target level).
    Let the maximum inventory gap $\bar{\delta}^n = \max_{j \in \mathcal{A}^n} \delta_j$.
    Let the demand set $\mathcal{D}^n = \emptyset$.
    while $\sum_{d \in \mathcal{D}^n} d \le \bar{\delta}^n$ do    ▷ while the cumulative demand has not exceeded $\bar{\delta}^n$
        Apply the policy $(\delta^n, S^n) := (\delta_{j^n}, \max(x^n, S_{j^n}^n))$.
        Append the realized sales into the demand set $\mathcal{D}^n$.
    end while
    Complete the current cycle: keep running $(\delta^n, S^n)$ policy until inventory drops below $S^n - \delta^n$.
    for $j \in \mathcal{A}^n$ do    ▷ SAAP and SGD
        Simulate the $(S_j, \delta_j)$ policy for one cycle using the collected demands in $\mathcal{D}^n$.
        Compute the simulated cycle pseudo cost $G_j^n$, cycle length $L_j^n$ and stochastic gradient $\tilde{\nabla}_j^n$.
        Update the target inventory level    ▷ Running the first-order optimization
            $S_j^{n+1} = \mathbf{Proj}_{[\delta_j, \beta]}\left(S_j^n - \eta_n \tilde{\nabla}_j^n\right), \quad \hat{G}_j^n = \hat{G}_j^{n-1} + G_j^n, \quad \hat{L}_j^n = \hat{L}_j^{n-1} + L_j^n.$
    end for
    Update and prune the active set    ▷ Running the zeroth-order optimization
        $\mathcal{A}^{n+1} = \left\{ j \in \mathcal{A}^n : \frac{\hat{G}_j^n}{\hat{L}_j^n} - \min_{j' \in \mathcal{A}^n} \frac{\hat{G}_{j'}^n}{\hat{L}_{j'}^n} \le \Delta^n \right\}.$
end for
```

implement the policy

$$\left(\max_{j \in \mathcal{A}^n} \delta_j, \max_{j \in \mathcal{A}^n} S_j^n \right) = \left(\bar{\delta}^n, S_{j^n}^n \right)$$

where j^n and $\bar{\delta}^n$ are given in (5.12) and (5.14), respectively, we would be able to simulate all other active policies. However, the problem is that the policy $(\bar{\delta}^n, S_{j^n}^n)$ itself may not be in the active set $\mathcal{A}^n$. In fact, we could implement $(\delta_{j^n}, S_{j^n}^n)$ at best (as it belongs to $\mathcal{A}^n$). Comparing the two policies $(\delta_{j^n}, S_{j^n}^n)$ and $(\bar{\delta}^n, S_{j^n}^n)$, there is a gap $\bar{\delta}^n - \delta_{j^n}$ that requires additional demand information.

Our solution is to extend the implementation of $(\delta_{j^n}, S_{j^n}^n)$ for potentially more than one complete cycles until the cumulative demand exceeds $\bar{\delta}^n$, so that the cumulative demand information

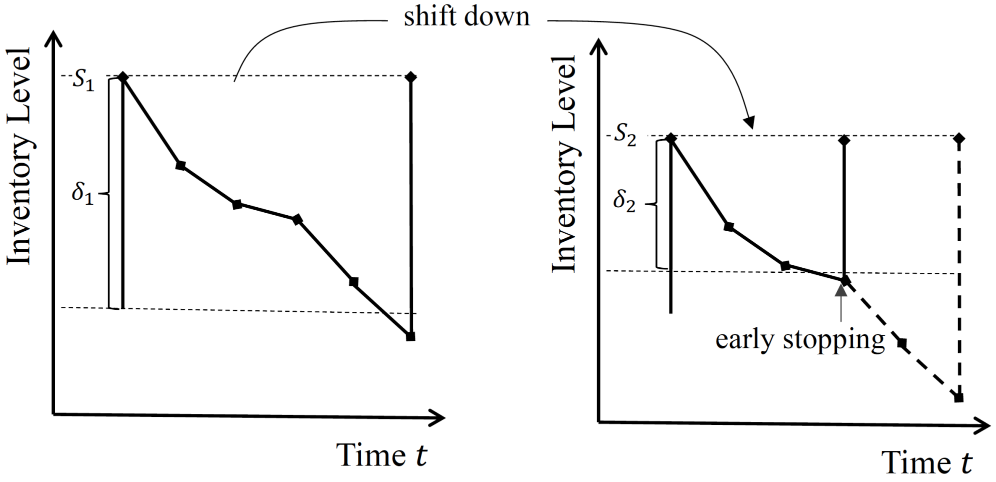


Figure 5.2: Simulation of another policy by “shifting down” the demands

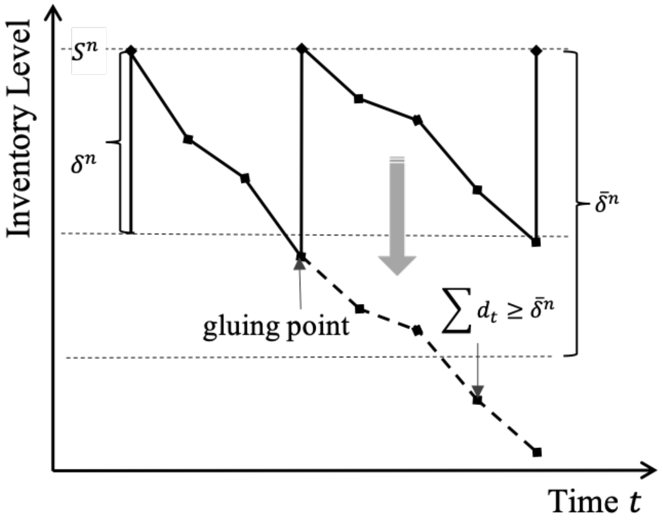


Figure 5.3: Simulation of another policy by “gluing” the Demands

(“glued” across cycles) becomes sufficient to simulate all the policies within the active set. Figure 5.3 gives an example that “glues” two cycles of demands obtained from running the policy $(\delta_{j^n}, S^n_{j^n})$ so that the cumulative demand exceeds the maximum inventory gap $\bar{\delta}^n$, thereby being able to simulate the policy $(\bar{\delta}^n, S^n_{j^n})$ and, hence, all policies within the active set.

It is also worth noting that the size of the active set $\mathcal{A}^n$ plays a key role in balancing the trade-off between *exploration* and *exploitation*. A larger active set $\mathcal{A}^n$ enables the decision maker to collect more information across more policies, whereas a smaller active set $\mathcal{A}^n$ focuses on the policies with sound hitherto empirical performance. In our performance analysis, we shall prove that our algorithm can achieve the optimal trade-off, by carefully choosing an adaptive confidence size Δ^n.

5.4 Performance Analysis of the (δ, S) Learning Algorithm

For ease of notation, we use the number of epochs N instead of the number of periods T to measure the regret, and we shall show that our (δ, S) algorithm achieves $O\left(\log N\sqrt{N}\right)$ regret, which clearly implies $O\left(\log T\sqrt{T}\right)$, since $N \leq T$. Note that Zhang et al. (2019) has established a lower bound $\Omega(\sqrt{T})$ for the repeated newsvendor problem (with no fixed cost and inventory carryover). Therefore, our regret bound is tight, up to a logarithmic factor.

Theorem 5.3 below states the main result of this paper.

Theorem 5.3 (Regret Bound)**.** *For any given integer $N \geq 1$, if we set the parameters $J = \lfloor\sqrt{N}\rfloor$, $\eta_n = \frac{\beta}{\xi\sqrt{n}}$, and $\Delta^n = \frac{2\theta\log(8N^2)}{\sqrt{n}}$, where the constants ξ and θ are given in* (5.28) *and* (5.32)*, respectively, then our (δ, S) learning algorithm has regret $\mathcal{R}_N = O\left(\log N\sqrt{N}\right)$.*

Note that Theorem 5.3 requires the prior knowledge of N, which may not be always available. However, we can readily apply the so-called *doubling trick* from online learning literature to remove this prior dependence on N to achieve an "anytime" algorithm. The idea is as follows.

1. We partition epochs into groups of exponentially increasing lengths.

2. We apply our algorithm for each group with parameters chosen according to the group length.

Then the new regret obtained for the anytime algorithm remains the same.

Theorem 5.4 (Regret Bound for Anytime Algorithm)**.** *We divide epochs into groups where the m^{th} ($m = 0, 1, 2, \ldots$) group contains epochs $\{2^m, \ldots, 2^{m+1} - 1\}$. For the m^{th} group, we apply the (δ, S) learning algorithm with parameters $J = \lfloor\sqrt{2^m}\rfloor$, $\eta_{2^m+n-1} = \frac{\beta}{\xi\sqrt{n}}$ and $\Delta^{2^m+n-1} = \frac{2\theta\log(2^{m+3})}{\sqrt{n}}$, where $1 \leq n \leq 2^m$ and the constants ξ and θ are given in* (5.28) *and* (5.32)*, respectively. Then our (δ, S) learning algorithm has regret $\mathcal{R}_N = O\left(\log N\sqrt{N}\right)$.*

Proof. Proof of Theorem 5.4. By Theorem 5.3, given the total number of epochs N, our (δ, S) learning algorithm achieves $O\left(\log N\sqrt{N}\right)$ regret. More explicitly, there exist some constants $\alpha_1, \alpha_2 \geq 0$,

$$\mathcal{R}_N \leq \alpha_1 \log N\sqrt{N} + \alpha_2. \tag{5.17}$$

Now, suppose the number of epoch N is unknown. Considering the first $m+1$ groups (where $m = 0, 1, 2\ldots$) contain $1 + 2 + 2^2 + \cdots + 2^m = 2^{m+1} - 1$ periods in total, if $M := \min\left\{m : 2^{m+1} - 1 \geq N\right\}$, then we have

$$\mathcal{R}_N \leq \sum_{m=0}^{M}\left(\alpha_1(\log 2)m\sqrt{2^m} + \alpha_2\right), \tag{5.18}$$

where the inequality is because, for the m^{th} group, we apply our algorithm with *known* epoch length 2^m, and therefore, following (5.17), the regret is less than $\alpha_1(\log 2)m\sqrt{2^m}+\alpha_2$. Since

$$\sum_{m=0}^{M} m\sqrt{2^m} \le \int_0^{M+1} m\sqrt{2^m}dm = \frac{4+2\left[(M+1)\log 2-2\right]\sqrt{2^{(M+1)}}}{(\log 2)^2}$$

and $N \ge 2^M$ (by the definition of M), then (5.18) implies $\mathcal{R}_N = O\left(\log N\sqrt{N}\right)$. □

5.4.1 Proof of Theorem 5.3

The remainder of this section is to establish our main result, i.e., Theorem 5.3.

We first formally define the *epoch length* and *epoch pseudo cost* with respect to a (δ, S) policy and a maximum inventory gap $\bar{\delta}$ as follows. Recall that $\{\tau_i\}_{i=1}^{\infty}$ are reorder periods defined in (5.3), which divides the planning horizon into cycles. We can group cycles into epochs such that the cumulative demand within an epoch exceeds $\bar{\delta}$. More precisely, the n^{th} epoch consists of the set of periods $\{\tau_{i_n}, \tau_{i_n}+1, \ldots, \tau_{i_{n+1}}-1\}$ where

$$\tau_{i_1} = \tau_1, \text{ and } \tau_{i_{n+1}} = \min\left\{\tau_i : i > i_n \text{ and } \sum_{t=\tau_{i_n}}^{\tau_i - 1} D_t \ge \bar{\delta}\right\}, \text{ for } n = 2, 3, \ldots$$

Extending the notation of the cycle pseudo cost and cycle length defined in (5.4) with an additional argument $\bar{\delta}$, we use $\tilde{L}_n^{(\delta,S,\bar{\delta})}$ and $\tilde{G}_n^{(\delta,S,\bar{\delta})}$ to denote the length and pseudo cost of the n^{th} epoch, respectively, where $\tilde{G}_n^{(\delta,S,\bar{\delta})}$ includes the newsvendor cost over $[\tau_{i_n}, \ldots, \tau_{i_{n+1}})$ and the ordering cost at $\tau_{i_{n+1}}$. For notational convenience, we use $\tilde{L}(\delta, S, \bar{\delta})$ and $\tilde{G}(\delta, S, \bar{\delta})$ to denote the (time-generic) epoch length and epoch pseudo cost, respectively.

It is evident that the inventory level process regenerates at the beginning of each epoch τ_{i_n} ($n = 1, 2, \ldots$). By Assumption 5, we know that $\mathbb{E}\left[\tilde{G}(\delta, S, \bar{\delta})\right] < \infty$, and then by apply the Renewal Reward Theorem (see Ross (1996)) and (5.8), we have

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \tilde{C}_t^{(\delta,S)}\right] = \frac{\tilde{G}(\delta,S,\bar{\delta})}{\tilde{L}(\delta,S,\bar{\delta})} = \frac{G(\delta,S)}{L(\delta,S)} = V(\delta,S). \tag{5.19}$$

We remark here that $\tilde{G}(\delta^n, S^n, \bar{\delta}^n)$ is not exactly describing the total cost occurred in the n^{th} epoch of our learning algorithm, because at the end of the n^{th} epoch, for making the transition to next epoch, we shall order up to S^{n+1} rather than S^n, which leads to a cost difference of size $c(S^{n+1} - S^n)$.

Without loss of generality, we assume $x_0 = 0$ and we order S^1 at the first period, which leads to an ordering cost $K+cS^1$ that is not included in any epoch. We can write the regret as follows.

$$
\begin{aligned}
\mathcal{R}_N &= \mathbb{E}\left[K + cS^1 + \sum_{n=1}^{N}\left(\tilde{G}(\delta^n, S^n, \bar{\delta}^n) + c(S^{n+1} - S^n)\right) - TV^*\right] \\
&= \sum_{n=1}^{N}\mathbb{E}\left[\tilde{G}(\delta^n, S^n, \bar{\delta}^n)\right] - \mathbb{E}[T]\, V^* + cS^{N+1} + K \\
&= \sum_{n=1}^{N}\frac{\mathbb{E}\left[\tilde{G}(\delta^n, S^n, \bar{\delta}^n)\right]}{\mathbb{E}\left[\tilde{L}(\delta^n, S^n, \bar{\delta}^n)\right]}\cdot\mathbb{E}\left[\tilde{L}(\delta^n, S^n, \bar{\delta}^n)\right] - \mathbb{E}\left[\sum_{n=1}^{N}\tilde{L}(\delta^n, S^n, \bar{\delta}^n)\right]V^* + cS^{N+1} + K \\
&= \mathbb{E}\left[\tilde{L}(S^n, \delta^n, \bar{\delta}^n)\right]\mathbb{E}\left[\sum_{n=1}^{N}\left(\frac{\mathbb{E}\left[\tilde{G}(\delta^n, S^n, \bar{\delta}^n)\right]}{\mathbb{E}\left[\tilde{L}(\delta^n, S^n, \bar{\delta}^n)\right]} - V^*\right)\right] + cS^{N+1} + K \\
&\le 2\mathbb{E}\left[\bar{L}\right]\cdot\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n) - V^*)\right] + c\beta + K, \qquad (5.20)
\end{aligned}
$$

where the last inequality follows from that $S^{N+1} \le \beta$ and (5.19) and

$$
\mathbb{E}\left[\tilde{L}\left(\delta^n, S^n, \bar{\delta}^n\right)\right] \le \mathbb{E}\left[2\bar{L}\right] \quad \text{for } n = 1, 2, \ldots, N. \qquad (5.21)
$$

Note that (5.21) holds true because, in the n^{th} epoch, we first wait until the cumulative demand exceed $\bar{\delta}^n$, and then complete the potentially incomplete cycle, i.e., $\mathbb{E}\left[\tilde{L}\left(\delta^n, S^n, \bar{\delta}^n\right)\right] \le \mathbb{E}\left[L(\bar{\delta}^n) + L(\delta^n)\right]$.

To obtain the regret bound, by (5.20), it suffices to bound the difference between $V(\delta^n, S^n)$ and V^*. To this end, we consider three intermediate policies bridging the original learning policy with the clairvoyant optimal policy, as shown in Figure 5.4.

1. *Learning Policy*: Apply the policy (δ^n, S^n) where $S^n = \max(x^n, S^n_{j^n})$ as in Algorithm 5.1.

2. *Bridging Policy I*: Apply the policy $(\delta^n, S^n_{j^n})$ whose long-run average pseudo cost is $V(\delta^n, S^n_{j^n})$.

3. *Bridging Policy II*: Apply the policy $(\delta^n, S^*(\delta^n))$ whose long-run average pseudo cost is

$$
V^*_{j^n} := V(\delta^n, S^*(\delta^n)), \qquad \text{where } S^*(\delta^n) = \arg\min_{s\in[\delta^n,\beta]} V(\delta^n, S).
$$

   That is, given any fixed δ^n, we apply the optimal order-up-to level $S^*(\delta^n)$.

4. *Bridging Policy III*: Apply the policy $(\delta_{j^*}, S^*(\delta_{j^*}))$ whose long-run average pseudo cost is $V^*_{j^*}$ where $j^* = \arg\min_{j\in\{1,\ldots,J\}} V^*_j$. That is, we apply the optimal discrete inventory gap δ_{j^*} and the corresponding optimal order-up-to level $S^*(\delta_{j^*})$.

5. *Clairvoyant Optimal Policy*: Apply the policy (δ^*, S^*).

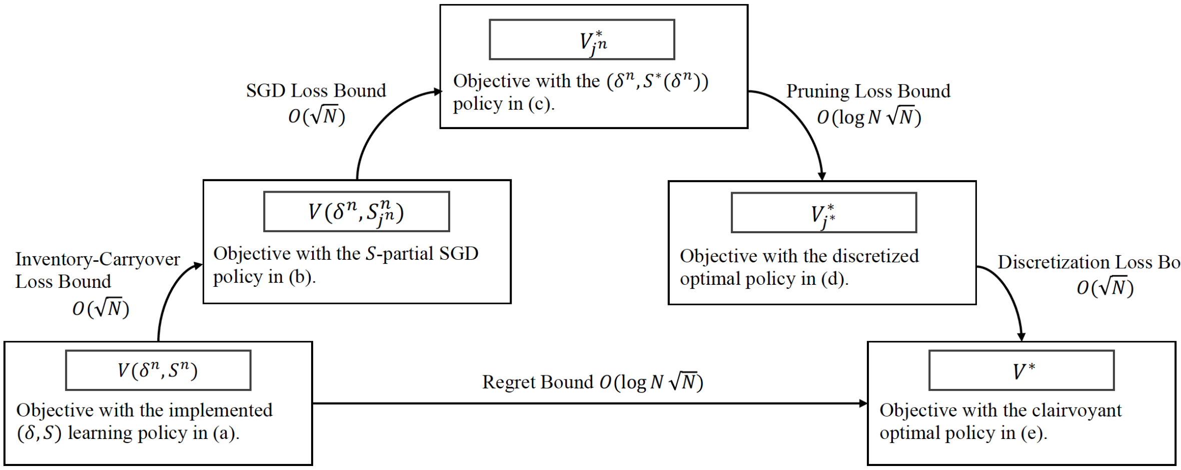

$V_{j^n}^*$
Objective with the $(\delta^n, S^*(\delta^n))$ policy in (c).
SGD Loss Bound $O(\sqrt{N})$
Pruning Loss Bound $O(\log N\sqrt{N})$
$V(\delta^n, S_{j^n}^n)$
Objective with the S-partial SGD policy in (b).
$V_{j^*}^*$
Objective with the discretized optimal policy in (d).
Inventory-Carryover Loss Bound $O(\sqrt{N})$
Discretization Loss Bound $O(\sqrt{N})$
$V(\delta^n, S^n)$
Objective with the implemented (δ, S) learning policy in (a).
Regret Bound $O(\log N\sqrt{N})$
V^*
Objective with the clairvoyant optimal policy in (e).

Figure 5.4: A series of bridging policies for the performance analysis

Given the above bridging policies, we decompose $\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n) - V^*)\right]$ in (5.20) as follows.

$$\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n) - V^*)\right] = \mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n) - V(\delta^n, S_{j^n}^n))\right] + \mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S_{j^n}^n) - V_{j^n}^*)\right]$$
$$+ \mathbb{E}\left[\sum_{n=1}^{N}\left(V_{j^n}^* - V_{j^*}^*\right)\right] + N(V_{j^*}^* - V^*). \tag{5.22}$$

Thus, to show Theorem 5.3, it suffices to show the following propositions.

Proposition 5.1 (Discretization Loss Bound)**.**

$$N(V_{j^*}^* - V^*) \leq O(\sqrt{N}). \tag{5.23}$$

Proposition 5.2 (Pruning Loss Bound)**.**

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V_{j^n}^* - V_{j^*}^*\right)\right] = O\left(\log N\sqrt{N}\right). \tag{5.24}$$

Proposition 5.3 (Inventory-Carryover Loss Bound)**.**

$$\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n) - V(\delta^n, S_{j^n}^n))\right] = O(\sqrt{N}). \tag{5.25}$$

Proposition 5.4 (SGD Loss Bound)**.**

$$\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n_{j^n}) - V^*_{j^n})\right] = O\left(\sqrt{N}\right). \tag{5.26}$$

Proof. Proof of Theorem 5.3. The result is an immediate consequence of Propositions 5.1–5.4. □

High-Level Intuitions: Let us provide high-level intuitions, before delving into the analysis.

1. Proposition 5.1 concerns the fourth term on the RHS of (5.22), which links the clairvoyant optimal policy with the bridging policy III. Since the bridging policy III is the optimal policy on the discrete grid $\{\delta_j\}_{j\in[J]}$, the loss of regret is due to the so-called *discretization loss*. Such a loss can be bounded by carefully choosing the grid size for discretization and also relying on the Lipschitz continuity of the objective function.

2. Proposition 5.2, which contains the *major innovation* of this paper, concerns the third term on the RHS of (5.22), which links the bridging policy III with the bridging policy II. Note that the bridging policy III gives us $V^*_{j^*}$ which is the optimal policy using the best δ_{j^*} on the grid $\{\delta_j\}_{j\in[J]}$, and the bridging policy II gives us $V^*_{j^n}$ which is the optimal policy using δ_{j^n} where $j^n \in \mathcal{A}_n$ is prescribed by our algorithm. Note that $\mathcal{A}_1 = [J]$ at the start, and the algorithm gradually removes candidate policies (based on their hitherto empirical performances) from the active set $\mathcal{A}_n$. It is critical that we ensure that the best δ_{j^*} is not being removed from the active set with an overwhelmingly high probability. To achieve this, we need to analyze the empirical estimator of the objective value, develop a new high probability bound for sub-exponential stochastic gradient descent (SGD), and also carefully choose the confidence size to update our active set in each iteration. Bounding such a loss is *central* to our regret analysis, and integrates the powers of SGD and bandit controls in a seamless and non-trivial fashion.

3. Proposition 5.3 concerns the first term on the RHS of (5.22), which links the bridging policy I with the original learning policy. The original learning policy sometimes cannot attain the desired target inventory level if the ending inventory carried over from the previous period is higher. Such a difference between the target level and the actual implemented level introduces an additional loss. We develop a new queueing system argument (that differs from existing ones) to bound the loss of not being able to immediately adjust to the desired target level.

4. Proposition 5.4 concerns the second term on the RHS of (5.22), which links the bridging policy II with the bridging policy I. The high-level idea is to ensure the prescribed target

level $S^n_{j^n}$ is approaching the optimal target level $S^*_{j^n}$ uniformly fast for all $\delta^n = \delta_{j^n}$, which can be achieved using our SAAP approach. The key is to be able to simulate all candidate policies within the active set $\mathcal{A}_n$ at each iteration, and therefore we can run SGD on all of them in a synchronized manner. Building upon the algorithmic construction, the analysis of this loss is standard.

We reiterate here that the previous approaches used in Huh et al. (2009); Huh and Rusmevichientong (2009); Zhang et al. (2018, 2019) rely heavily on the convexity of the objectives, and therefore *need not* maintain and adaptively prune an active set of all favorable policies. Taking into account of fixed cost, just as every other paper in inventory theory (be it learning or not), changes the landscape of the problem and spurs a new methodological development.

5.4.2 Proof of Proposition 5.1 – Bounding the Discretization Loss

We first bound the last term on the RHS of (5.22). The intuition is simple. That is, this term captures the difference between the "discrete" optimal long-run average pseudo cost on the grid $\{\delta_j\}_{j\in[J]}$ and the true optimal long-run average pseudo cost, which can be bounded by choosing the grid size $J = \lfloor\sqrt{N}\rfloor$ and using the Lipschitz continuity of $V^*(\delta)$ in (5.11).

Proof. Proof of Proposition 5.1. Note that we initialize the active set $\mathcal{A}^1$ with size $J = \lfloor\sqrt{N}\rfloor$. This implies that if we choose δ_k on the grid $\{\delta_j\}_{j\in[J]}$ that is closest to the clairvoyant optimal δ^*, i.e.,

$$k = \arg\min_{j\in[J]} |\delta^* - \delta_j|,$$

then $|\delta^* - \delta_k| \leq \beta/\lfloor\sqrt{N}\rfloor$. Furthermore, by the Lipschitz continuity of $V^*(\delta)$ in Theorem 5.2, we have

$$V^*_{j^*} - V^* = V^*(\delta_{j^*}) - V^*(\delta^*) \leq V^*(\delta_k) - V^*(\delta^*) \leq O(1/\sqrt{N}),$$

which leads the desired result. □

5.4.3 Proof of Proposition 5.2 – Bounding the Pruning Loss

For each epoch n, the active set $\mathcal{A}_n$ contains the indices of polices with high hitherto empirical performances. Recall the cumulative empirical $\hat{G}^n_j$ and $\hat{L}^n_j$ defined in (5.15). Let

$$\hat{V}^n_j := \hat{V}^n\left(\delta_j\right) := \frac{\hat{G}^n_j}{\hat{L}^n_j} \tag{5.27}$$

denote our n-step estimator for the optimal long-run average pseudo cost with the inventory gap δ_j in the n^{th} epoch. First, we shall argue that this empirical estimator $\hat{V}_j^n$ is a good approximation for $V_j^* := V^*\left(\delta_j\right)$. To achieve that, we will need the following results.

Lemma 4 (Sub-Exponentials)**.** *There exist positive constants v, b and ξ such that, for any feasible (δ, S) policy, the following statements hold.*

1. *The corresponding cycle length L, cycle cost G and stochastic gradient $\tilde{\nabla}$ defined in* (5.10) *are sub-exponential with parameters* (v, b) *(see Definition 7).*

2. *The second moment of the stochastic gradient $\tilde{\nabla}$ is bounded by ξ^2, i.e.,*

$$\mathbb{E}\left[\tilde{\nabla}^2\right] \leq \xi^2, \quad \textit{where} \quad \xi := \max(p, c, h)\sqrt{\mathbb{E}\left[\bar{L}^2\right]}. \tag{5.28}$$

Proof. **Step I**. We shall show $\bar{L} = L(\beta)$ is sub-exponential by arguing that $\mathbb{P}[L(\beta) > t]$ decays sub-exponentially as t increases. Let F be the cdf of demand D_t. Define $U_t := F(D_t)$, so U_t follows the uniform distribution on $[0,1]$. Since D is positive and ρ bounds the pdf of D_t, we have $F(0) = 0$ and $\|F'\|_\infty \leq \rho$, which implies, for any $x \geq 0$, we have $F(x) \leq \rho x$. Hence, from the definition of U, $U_t \leq \rho D_t$. Without loss of generality, let t be an even positive integer. Consider

$$\begin{aligned}
\mathbb{P}[L(\beta) \geq t] = \mathbb{P}\left[\sum_{i=1}^{t} D_i < \beta\right] \leq \mathbb{P}\left[\sum_{i=1}^{t} U_t < \beta\rho\right] &= \frac{1}{t!}\sum_{k=0}^{\lfloor \beta\rho \rfloor}(-1)^k \binom{t}{k}(\beta\rho - k)^t \\
&= \frac{1}{t!}\sum_{k=0}^{\lfloor \beta\rho \rfloor}(-1)^k \frac{t!}{k!(t-k)!}(\beta\rho - k)^t \leq \frac{1}{(\frac{t}{2}!)^2}\sum_{k=0}^{\lfloor \beta\rho \rfloor}(\beta\rho - k)^t \\
&\leq \frac{1}{(\frac{t}{2}!)^2}(\beta\rho + 1)(\beta\rho)^t \leq e^4(\beta\rho + 1)\left(\frac{e}{\beta\rho}\right)^{-t}
\end{aligned}$$

where the second equality holds because the sum of uniform random variables $\sum_{i=1}^{t} U_t$ follows the Irwin-Hall distribution with parameter t, and the last inequality is because $1/(\frac{t}{2}!)^2 \leq e^{4-t}$ for all $t \geq 0$. Thus, we conclude that $\bar{L}$ is sub-exponential.

Step II. We shall show that there exists (v_1, b_1) such that, for any policy (δ, S), its cycle length L is (v_1, b_1) sub-exponential. Since $\bar{L}$ is sub-exponential, by the equivalent characterization in Theorem 5.7, there is a positive constant b such that $\mathbb{E}\left[\exp\left(\lambda\left(\bar{L} + \mathbb{E}\left[\bar{L}\right]\right)\right)\right] < \infty$ for any $\lambda \leq 1/b$. By Taylor's expansion,

$$\mathbb{E}\left[\exp\left(\lambda\left(\bar{L} + \mathbb{E}\left[\bar{L}\right]\right)\right)\right] = 1 + \lambda\mathbb{E}\left[\left(\bar{L} + \mathbb{E}\left[\bar{L}\right]\right)\right] + \frac{\lambda^2\mathbb{E}\left[\left(\bar{L} + \mathbb{E}\left[\bar{L}\right]\right)^2\right]}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[\left(\bar{L} + \mathbb{E}\left[\bar{L}\right]\right)^k\right]}{k!},$$

which implies that

$$\frac{\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(\bar{L}+\mathbb{E}[\bar{L}])^k\right]}{k!}}{\lambda^2}=\frac{\mathbb{E}\left[\exp\left(\lambda\left(\bar{L}+\mathbb{E}\left[\bar{L}\right]\right)\right)\right]-1-\lambda\mathbb{E}\left[\left(\bar{L}-\mathbb{E}\left[\bar{L}\right]\right)\right]+\frac{\lambda^2\mathbb{E}\left[(\bar{L}+\mathbb{E}[\bar{L}])^2\right]}{2}}{\lambda^2}. \tag{5.29}$$

By L’Hôspital’s rule, we have that (5.29) converges to 0 as $\lambda \to 0$, which implies that

$$\frac{\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(\bar{L}+\mathbb{E}[\bar{L}])^k\right]}{k!}}{\lambda^2}=o\left(\lambda^2\right). \tag{5.30}$$

Now, consider

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda(L-\mathbb{E}L))\right] &= 1+\frac{\lambda^2\mathbb{E}\left[(L-\mathbb{E}L)^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(L-\mathbb{E}L)^k\right]}{k!}\\
&\le 1+\frac{\lambda^2\mathbb{E}\left[L^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(L+\mathbb{E}\left[L\right])^k\right]}{k!}\\
&\le 1+\frac{\lambda^2\mathbb{E}\left[\bar{L}^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[\left(\bar{L}+\mathbb{E}\left[\bar{L}\right]\right)^k\right]}{k!}\\
&\le 1+\frac{\lambda^2\mathbb{E}\left[\bar{L}^2\right]}{2}+o\left(\lambda^2\right),
\end{aligned}$$

where the last inequality follows from (5.30).

On the other hand, for all $\nu > 0$, we have

$$\exp\left(\frac{\lambda^2\nu^2}{2}\right)=1+\frac{\lambda^2\nu^2}{2}+o\left(\lambda^2\right).$$

Note, since $\bar{L}$ is sub-exponential, its second moment $\mathbb{E}\left[\bar{L}^2\right]<\infty$. If we choose ν_1 such that $\nu_1^2 > \mathbb{E}\left[\bar{L}^2\right]$, there exists $b_1 > 0$ such that

$$\mathbb{E}\left[\exp(\lambda(L-\mathbb{E}L))\right]\le\exp\left(\frac{\lambda^2\nu_1^2}{2}\right),\text{ for all } |\lambda|\le\frac{1}{b_1},$$

which shows that L is (ν_1, b_1)-sub-exponential.

Step III. We shall apply similar arguments to show that there exists (ν_2, b_2) such that, for any policy (δ, S), its cycle cost G is (v_2, b_2) sub-exponential.

$$
\begin{aligned}
\mathbb{E}\left[\exp(\lambda(G-\mathbb{E}G))\right] &= 1+\frac{\lambda^2\mathbb{E}\left[(G-\mathbb{E}G)^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(G-\mathbb{E}G)^k\right]}{k!} \\
&= 1+\frac{\lambda^2\mathbb{E}\left[G^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(\gamma L+\gamma\mathbb{E}L)^k\right]}{k!} \\
&\leq 1+\lambda\frac{\gamma^2\mathbb{E}\left[L^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[\left(\gamma\bar{L}+\gamma\mathbb{E}\bar{L}\right)^k\right]}{k!} \\
&\leq 1+\lambda\frac{\gamma^2\mathbb{E}\left[\bar{L}^2\right]}{2}+o\left(\lambda^2\right).
\end{aligned}
$$

Therefore, if we choose v_2 such that $v_2^2 > \gamma^2\mathbb{E}\left[\bar{L}^2\right]$, there exists $b_2 > 0$ such that

$$
\mathbb{E}\left[\exp(\lambda(G-\mathbb{E}G))\right] \leq \exp\left(\frac{\lambda^2 v_2^2}{2}\right), \text{ for all } |\lambda| \leq \frac{1}{b_2},
$$

which shows that G is also (v_2, b_2)-sub-exponential.

Step IV. We shall show that there exists (v_3, b_3) such that, for any policy (δ, S), the stochastic gradient $\tilde{\nabla}$ defined in (5.10) is (v_3, b_3) sub-exponential, and, consequently, has a bounded second moment. By the definition $\tilde{\nabla}$ in (5.10), we have $|\tilde{\nabla}| \leq \max(c,h,p)L$. Again, by applying similar arguments, we have

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\left(\tilde{\nabla}-\mathbb{E}\tilde{\nabla}\right)\right)\right] &= 1+\frac{\lambda^2\mathbb{E}\left[\left(\tilde{\nabla}-\mathbb{E}\left[\tilde{\nabla}\right]\right)^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[\left(\tilde{\nabla}-\mathbb{E}\left[\tilde{\nabla}\right]\right)^k\right]}{k!} \\
&\leq 1+\frac{\lambda^2\mathbb{E}\left[\tilde{\nabla}^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[(\max(c,h,p)L+\mathbb{E}\left[\max(c,h,p)L\right])^k\right]}{k!} \\
&\leq 1+\frac{\lambda^2(\max(c,h,p))^2\mathbb{E}\left[\bar{L}^2\right]}{2}+\sum_{k=3}^{\infty}\frac{\lambda^k(\max(c,h,p))^k\mathbb{E}\left[\left(\bar{L}+\mathbb{E}\left[\bar{L}\right]\right)^k\right]}{k!} \\
&\leq 1+\frac{\lambda^2(\max(c,h,p))^2\mathbb{E}\left[\bar{L}^2\right]}{2}+o\left(\lambda^2\right).
\end{aligned}
$$

Therefore, if we choose v_3 such that $v_3^2 > \max(h,p,c)^2\mathbb{E}\left[\bar{L}^2\right]$, there exists $b_3 > 0$ such that

$$
\mathbb{E}\left[\exp\left(\lambda\left(\tilde{\nabla}-\mathbb{E}\tilde{\nabla}\right)\right)\right] \leq \exp\left(\frac{\lambda^2 v_3^2}{2}\right), \text{ for all } |\lambda| \leq \frac{1}{b_3},
$$

which shows that $\tilde{\nabla}$ is (v_3, b_3) sub-exponential. Since $|\tilde{\nabla}| \leq \max(p,c,h)\bar{L}$, we have (5.28). To complete the proof, we choose $v = \max(v_1, v_2, v_3)$ and $b = \max(b_1, b_2, b_3)$.

□

At the n^{th} epoch, for each index j of inventory gap δ_j that remains in the active set $\mathcal{A}_n$, we want to bound the *sub-optimality gap* between the empirical estimator $\hat{V}_j^n$ and the estimand V_j^* with a high probability, which is formally stated in Lemma 5 below.

However, to achieve the desired results in Lemma 5, we shall develop a new high probability regret bound for the general *sub-exponential* SGD algorithm in Theorem 5.5 below. Since Theorem 5.5 is very useful and of independent interest outside the context of this paper, we state and prove the results with a general convex function f with argument z.

Theorem 5.5 (High Probability Regret Bound for Sub-Exponential SGD)**.** *Let $\{z_i\}_{i=1}^n$ be a sequence generated by the projected stochastic gradient descent algorithm with respect to a convex function $f(\cdot)$ with a domain $\mathcal{K}$, i.e.,*

$$z_1 \in \mathcal{K} \quad \textit{and} \quad z_{i+1} = \mathbf{Proj}_{\mathcal{K}}\left[z_i - \eta_i \tilde{\nabla}_i\right] \quad \textit{for } i = 1, \ldots, n-1,$$

where $\tilde{\nabla}_i$ is a stochastic gradient of f at x_i and η_i is the step size in the i^{th} iteration.

We make the following assumptions:

1. *The diameter of function domain $\mathcal{K}$ is bounded by β. i.e., $\sup_{z_1, z_2 \in \mathcal{K}} \|z_1 - z_2\| \leq \beta$.*

2. *For $i = 1, 2, \ldots, n-1$, conditional on z_i, the stochastic gradient $\tilde{\nabla}_i$ is (v, b)-sub-exponential random vector (see Definition 8) with the second moment bounded by a positive constant ξ^2*
.

Then, if choosing the step size $\eta_i = \frac{\beta}{\xi\sqrt{i}}$, we have, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}[f(z_i) - f(z^*)] \leq \max\left\{\sqrt{\frac{2\beta^2 v^2 \log(1/\delta)}{n}}, \frac{2b\log(1/\delta)}{n}\right\} + \frac{3\beta\xi}{2\sqrt{n}},$$

where $z^* = \arg\min_{z \in \mathcal{K}} f(z)$.

Proof. Proof. Since f is convex, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[f(z_i) - f(z^*)] \leq & \frac{1}{n}\sum_{i=1}^{n}\langle \nabla f(z_i), z_i - z^* \rangle \\ = & \frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla f(z_i) - \tilde{\nabla}_i, z_i - z^* \right\rangle + \frac{1}{n}\sum_{i=1}^{n}\left\langle \tilde{\nabla}_i, z_i - z^* \right\rangle. \end{aligned} \tag{5.31}$$

Step I: We shall show that the second term on the RHS of (5.31)

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle\tilde{\nabla}_i, z_i - z^*\right\rangle \leq \frac{3\xi\beta}{2\sqrt{n}} \quad \text{almost surely.}$$

Note that

$$\left\|z_{i+1} - z^*\right\|^2 = \left\|\mathbf{Proj}_{\mathcal{K}}\left(z_i - \eta_i\tilde{\nabla}_i\right) - z^*\right\|^2 \leq \left\|z_i - \eta_i\tilde{\nabla}_i - z^*\right\|^2.$$

Hence,

$$\left\|z_{i+1} - z^*\right\|^2 \leq \left\|z_i - z^*\right\|^2 + \eta_i^2\left\|\tilde{\nabla}_i\right\|^2 - 2\eta_i\left\langle\tilde{\nabla}_i, z_i - z^*\right\rangle,$$

which implies that

$$2\left\langle\tilde{\nabla}_i, z_i - z^*\right\rangle \leq \frac{\|z_i - z^*\|^2 - \|z_{i+1} - z^*\|^2}{\eta_i} + \eta_i\left\|\tilde{\nabla}_i\right\|^2.$$

Therefore,

$$\begin{aligned}
2\sum_{i=1}^{n}\left\langle\tilde{\nabla}_i, z_i - z^*\right\rangle &\leq \sum_{i=1}^{n}\left(\frac{\|z_i - z^*\|^2 - \|z_{i+1} - z^*\|^2}{\eta_i} + \eta_i\left\|\tilde{\nabla}_i\right\|^2\right) \\
&\leq \sum_{i=1}^{n}\left(\left\|z_i - z^*\right\|^2\left(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}}\right) + \eta_i\xi^2\right) \\
&\leq \beta^2\frac{1}{\eta_n} + \sum_{i=1}^{n}\eta_i\xi^2 \\
&\leq 3\beta\xi\sqrt{n},
\end{aligned}$$

where $\frac{1}{\eta_0} := 0$, and the last inequality follows from that $\eta_i = \frac{\beta}{\xi\sqrt{i}}$ and $\sum_{i=1}^{n} 1/\sqrt{i} \leq 2\sqrt{n}$.

Step II: We shall show that the first term on the RHS of (5.31)

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle\nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\right\rangle \leq \max\left\{\sqrt{\frac{2\beta^2\nu^2\log(1/\delta)}{n}}, \frac{2b\log(1/\delta)}{n}\right\}.$$

with probability at least $1-\delta$. Define the filtration $\mathcal{F}_i = \sigma\left(\tilde{\nabla}_1, \ldots, \tilde{\nabla}_i\right)$. Then we have

$$\mathbb{E}\left[\left\langle\nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\right\rangle|\mathcal{F}_{i-1}\right] = \mathbb{E}\left[\left\langle\nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\right\rangle|z_i\right] = 0,$$

which shows that $\{\langle\nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\rangle, \mathcal{F}_i\}_i$ is a martingale difference sequence (see Definition 6). Since $\tilde{\nabla}_i$ is (ν, b)-sub-exponential, we have that $\langle\nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\rangle$ is $(\beta\nu, b)$-sub-exponential.

Thus, by Azuma's inequality stated in Theorem 5.8, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\right\rangle \geq t\right] \leq \begin{cases} e^{-\frac{nt^2}{2\beta^2\nu^2}} & \text{for } 0 \leq t \leq \frac{\beta^2\nu^2}{b} \\ e^{-\frac{nt}{2b}} & \text{for } t > \frac{\beta^2\nu^2}{b} \end{cases} \leq \max\left\{e^{-\frac{nt}{2b}}, e^{-\frac{nt^2}{2\beta^2\nu^2}}\right\}$$

which implies that

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla f(z_i) - \tilde{\nabla}_i, z_i - z^*\right\rangle \leq \max\left\{\sqrt{\frac{2\beta^2\nu^2\log(1/\delta)}{n}}, \frac{2b\log(1/\delta)}{n}\right\}$$

with probability at least $1-\delta$. □

Lemma 5 below quantifies the estimation error between the empirical estimator $\hat{V}_j^n$ and the estimand V_j^* by running the stochastic gradient descent on S for the inventory gap δ_j.

Lemma 5 (Estimation Confidence Bound). *For any $\kappa \in (0,1)$, we have, with probability at least $1-\kappa$,*

$$\left|\hat{V}_j^n - V_j^*\right| \leq \frac{\theta\log(8/\kappa)}{\sqrt{n}},$$

where

$$\theta := (2\gamma+4)\max(\nu, b) + 1.5\beta\xi. \tag{5.32}$$

Proof. Proof. Recall that the estimand and the empirical estimator are given by

$$V_j^* = V(\delta_j, S_j^*) = \frac{\min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]}{\mathbb{E}\left[L\left(\delta_j\right)\right]}, \quad \text{and} \quad \hat{V}_j^n = \hat{V}^n(\delta_j) = \frac{\hat{G}_j^n}{\hat{L}_j^n}.$$

Note that $\mathbb{E}\left[G\left(\delta_j, S\right)\right]$ is convex with respect to S for each $j \in [J]$.

Step I: Bound the approximation error of the average cycle length $\left|\frac{\hat{L}_j^n}{n} - \mathbb{E}\left[L_j\right]\right|$. By Lemma 4, we have that L_j is (ν, b)-sub-exponential. Then by Corollary 5.9 of Azuma's inequality, we have with probability of at least $1-\kappa$,

$$\left|\frac{\hat{L}_j^n}{n} - \mathbb{E}\left[L\left(\delta_j\right)\right]\right| \leq \max\left(\sqrt{\frac{2\nu^2\log(2/\kappa)}{n}}, \frac{2b\log(2/\kappa)}{n}\right). \tag{5.33}$$

Step II: Bound the approximation error of the average cycle pseudo cost $\left|\frac{\hat{G}_j^n}{n} - \min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]\right|$.

$$\begin{aligned}
\left|\frac{\hat{G}_j^n}{n} - \min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]\right| &= \left|\frac{G\left(\delta_j, S_j^1\right) + \cdots + G\left(\delta_j, S_j^n\right)}{n} - \min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]\right| \\
&\leq \left|\frac{G\left(\delta_j, S_j^1\right) + \cdots + G\left(\delta_j, S_j^n\right)}{n} - \frac{\mathbb{E}\left[G\left(\delta_j, S_j^1\right)\right] + \cdots + \mathbb{E}\left[G\left(\delta_j, S_j^n\right)\right]}{n}\right| \\
&\quad + \left|\frac{\mathbb{E}\left[G\left(\delta_j, S_j^1\right)\right] + \cdots + \mathbb{E}\left[G\left(\delta_j, S_j^n\right)\right]}{n} - \min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]\right|. \quad (5.34)
\end{aligned}$$

To bound the second term on the RHS of (5.34), we have the following observations.

1. $\mathbb{E}[G(\delta_j, \cdot)]$ is convex by Theorem 5.2.
2. The sequence $\{S_j^k\}_{k=1}^n$ is generated by running stochastic gradient descent on $\mathbb{E}[G(\delta_j, \cdot)]$.
3. The diameter of $\mathbb{E}[G(\delta_j, \cdot)]$'s domain is bounded by β by Assumption 5.
4. The second moment of the stochastic gradient $\tilde{\nabla}$ is bounded by ξ^2 by Lemma 4.
5. The stochastic gradient $\tilde{\nabla}$ is (ν, b)-sub-exponential by Lemma 4.

Therefore, by Theorem 5.5, we have with probability of at least $1 - \kappa$,

$$\begin{aligned}
&\left|\frac{\mathbb{E}\left[G\left(\delta_j, S_j^1\right)\right] + \cdots + \mathbb{E}\left[G\left(\delta_j, S_j^n\right)\right]}{n} - \min_S \mathbb{E}\left[G\left(\delta_j, S\right)\right]\right| \\
&\leq \max\left\{\sqrt{\frac{2\beta^2\nu^2 \log(1/\kappa)}{n}}, \frac{2b\log(1/\kappa)}{n}\right\} + \frac{3\beta\xi}{2\sqrt{n}}.
\end{aligned}$$

To bound the first term on the RHS of (5.34), we note that

$$\left\{G\left(\delta_j, S_j^k\right) - \mathbb{E}\left[G\left(\delta_j, S_j^k\right)\right]\right\}_{k=1}^n$$

is a martingale difference sequence and, by Lemma 4, each term in the sequence is (ν, b)-sub-exponential. Then, by Azuma's inequality in Theorem 5.8, we have with probability at least $1 - \kappa$,

$$\begin{aligned}
&\left|\frac{G\left(\delta_j, S_1\right) + \cdots + G\left(\delta_j, S^n\right)}{n} - \frac{\mathbb{E}\left[G\left(\delta_j, S_1\right)\right] + \cdots + \mathbb{E}\left[G\left(\delta_j, S^n\right)\right]}{n}\right| \\
&\leq \max\left(\sqrt{\frac{2\nu^2 \log(2/\kappa)}{n}}, \frac{2b\log(2/\kappa)}{n}\right).
\end{aligned}$$

Combining the above two bounds for the RHS of (5.34), we have with probability at least $1-\kappa$,

$$\begin{aligned}\left|\frac{\hat{G}_j^n}{n}-\min_S \mathbb{E}\left[G\left(\delta_j,S\right)\right]\right| \leq &\max\left(\sqrt{\frac{2\nu^2\log(4/\kappa)}{n}},\frac{2b\log(4/\kappa)}{n}\right)\\ &+\max\left(\sqrt{\frac{2\beta^2\nu^2\log(2/\kappa)}{n}},\frac{2b\log(2/\kappa)}{n}\right)+\frac{3\beta\xi}{2\sqrt{n}}\\ \leq &2\max\left(\sqrt{\frac{2\beta^2\nu^2\log(4/\kappa)}{n}},\frac{2b\log(4/\kappa)}{n}\right)+\frac{3\beta\xi}{2\sqrt{n}}.\end{aligned} \tag{5.35}$$

<u>**Step III**</u>. Bound the estimation error $\left|\hat{V}_j^n-V_j^*\right|$. For better readability, we slight abuse the notation to define

$$G_1 := \frac{\hat{G}_j^n}{n},\ L_1 := \frac{\hat{L}_j^n}{n},\ G_2 := \min_S \mathbb{E}\left[G\left(\delta_j,S\right)\right],\ L_2 := \mathbb{E}\left[L\left(\delta_j\right)\right].$$

$\left|\hat{V}_j^n-V_j^*\right|$ can be decomposed as follows.

$$\left|\hat{V}_j^n-V_j^*\right| = \left|\frac{G_1}{L_1}-\frac{G_2}{L_2}\right| = \left|\frac{G_1}{L_1}-\frac{G_1}{L_2}+\frac{G_1}{L_2}-\frac{G_2}{L_2}\right| \leq \frac{1}{L_1}\left|G_1-G_2\right|+\left|\frac{G_2}{L_1L_2}(L_1-L_2)\right|.$$

Considering $L_1 \geq 1$ and $|G_2/L_2| \leq \gamma$, we have

$$\left|\hat{V}_j^n-V_j^*\right| \leq \left|\frac{\hat{G}_j^n}{n}-\min_S \mathbb{E}\left[G\left(\delta_j,S\right)\right]\right|+\gamma\left|\frac{\hat{L}_j^n}{n}-\mathbb{E}\left[L\left(\delta_j\right)\right]\right|.$$

Then, by (5.33) and (5.35), we have with probability at least $1-\kappa$,

$$\begin{aligned}\left|\hat{V}_j^n-V_j^*\right| \leq &2\max\left(\sqrt{\frac{2\nu^2\log(8/\kappa)}{n}},\frac{2b\log(8/\kappa)}{n}\right)+\frac{3\beta\xi}{2\sqrt{n}}+\gamma\max\left(\sqrt{\frac{2\nu^2\log(4/\kappa)}{n}},\frac{2b\log(4/\kappa)}{n}\right)\\ \leq &\frac{\theta\log(8/\kappa)}{\sqrt{n}},\end{aligned}$$

where θ is defined in (5.32). This completes the proof. □

Now, we are ready to prove Proposition 5.2. This proof spells out one "major backbone" of this paper that *simultaneously* bound the sub-optimality loss from running a sub-exponential SGD and the bandit loss from pruning the active set. One pivotal step is to ensure that conditioning on some overwhelming high probability event, the optimal policy on the grid does not get "removed" while updating the active set in each iteration. This guarantees our pruning accuracy. Moreover, via our developed high probability sub-exponential SGD bound, the probability of the complement event can also be bounded, resulting in a controlled loss on the complement event.

Proof. Proof of Proposition 5.2. Recall that the choice of Δ^n from Theorem 5.3 is given by

$$\Delta^n = \frac{2\theta \log(8N^2)}{\sqrt{n}}, \quad \text{where } \theta \text{ is given in (5.32).}$$

By Lemma 5, we have

$$\mathbb{P}\left[\left|\hat{V}_j^n - V_j^*\right| \geq \Delta^n/2\right] \leq \frac{1}{N^2}.$$

Define the event A and its complement A^c by

$$A = \left\{ \text{ for all } j \in [J], n \in [N] \text{ we have } \left|\hat{V}_j^n - V_j^*\right| < \Delta^n/2 \right\}.$$
$$A^c = \left\{ \text{ there exists some } j \in [J], n \in [N] \text{ such that } \left|\hat{V}_j^n - V_j^*\right| \geq \Delta^n/2 \right\}.$$

Since $J = \lfloor \sqrt{N} \rfloor$ and applying the union bound, we have

$$\mathbb{P}[A^c] \leq JN\left(\frac{1}{N^2}\right) \leq \frac{1}{\sqrt{N}} \quad \text{and} \quad \mathbb{P}[A] \geq 1 - \frac{1}{\sqrt{N}}.$$

Define $j^* = \arg\min_j V_j^*$, i.e., the optimal index on the grid.

Now we condition on the event A. Then for any $j \in [J]$ and $n \in [N]$, we have

$$\hat{V}_{j^*}^n - \hat{V}_j^n \leq \hat{V}_{j^*}^n - V_{j^*}^* + V_j^* - \hat{V}_j^n \leq \Delta^n,$$

which implies that

$$\hat{V}_{j^*}^n - \min_j \hat{V}_j^n \leq \Delta^n.$$

Then comparing with the updating rule on our active set (5.16), we conclude that j^* will always remain in the active set of every iteration and never leave. Since $j^n \in \mathcal{A}^n$ which implies j^n is not “removed” from the active set in the $(n-1)^{th}$ iteration, we have

$$\hat{V}_{j^n}^{n-1} - \hat{V}_{j^*}^{n-1} \leq \hat{V}_{j^n}^{n-1} - \min_{j \in \mathcal{A}^{n-1}} \hat{V}_j^{n-1} \leq \Delta^{n-1},$$

where the second inequality follows from our rule (5.16). Therefore, conditional on the event A,

$$\begin{aligned} V_{j^n}^* - V_{j^*}^* &= \left(V_{j^n}^* - \hat{V}_{j^n}^{n-1}\right) + \left(\hat{V}_{j^n}^{n-1} - \hat{V}_{j^*}^{n-1}\right) + \left(\hat{V}_{j^*}^{n-1} - V_{j^*}^*\right) \\ &\leq \frac{1}{2}\Delta^{n-1} + \Delta^{n-1} + \frac{1}{2}\Delta^{n-1} = 2\Delta^{n-1}, \end{aligned}$$

where $\Delta^0 := \gamma$.

Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}\left(V^*_{j^n}-V^*_{j^*}\right)\right] &= \mathbb{E}\left[\sum_{n=1}^{N}\left(V^*_{j^n}-V^*_{j^*}\right)|A\right]\mathbb{P}[A]+\mathbb{E}\left[\sum_{n=1}^{N}\left(V^*_{j^n}-V^*_{j^*}\right)|A^c\right]\mathbb{P}[A^c] \\
&\leq \mathbb{E}\left[\sum_{n=1}^{N}\left(V^*_{j^n}-V^*_{j^*}\right)|A\right]+\gamma\sqrt{N} \leq \sum_{n=1}^{N} 2\Delta^{n-1}+\gamma\sqrt{N} = O(\log N\sqrt{N}),
\end{aligned}$$

where the last equality holds by plugging in $\Delta^n = \frac{2\theta\log(8N^2)}{\sqrt{n}}$. □

5.4.4 Proof of Proposition 5.3 – Bounding the Inventory-Carryover Loss

Consider the case that the implemented policy in the current epoch n is (δ^n_i, S^n_i) and in the next epoch $n+1$ is $(\delta^{n+1}_j, S^{n+1}_j)$. If $S^n_i - \delta^n_i > S^{n+1}_j$, it is possible that the beginning epoch inventory level $x^{n+1} > S^{n+1}_j$. Then the algorithm cannot immediately adjust to the target level S^{n+1}_j, and is therefore forced to take x^{n+1} instead, which introduces a loss in (5.24).

To bound the loss due to such a positive inventory carryover, we model the inventory process as a **G/G/1** queue. We first consider a queueing system where in period n, Z_n is the queueing length, A_n is the inter-arrival time, and B_n is the service time. In our application, the queueing length corresponds to the inventory level, the inter-arrival time A_n corresponds to the adjustment due to a fixed inventory control policy, and the service time D_n corresponds to the random demand. We have the following lemma to bound the total waiting time.

Lemma 6 (Bound on Total Waiting Time)**.** *Given a queueing system, $Z_1 = 0$, and, for $n = 1,\ldots,N-1$,*

$$Z_{n+1} = \mathbf{Proj}_{[0,\beta]}(Z_n + A_n - D_n),$$

where β is the upper bound on the queueing length, the inter-arrival times $\{A_n\}_n$ satisfies $\sum_{n=1}^N A_n = O(\sqrt{N})$, and service times $\{D_n\}_n$ are i.i.d. positive random variables, then we have

$$\mathbb{E}\left[\sum_{n=1}^{N} Z_n\right] = O(\sqrt{N}).$$

Proof. Proof. Note that $\sum_{n=1}^N Z_n$ is the total waiting time for all customers in the queueing system. Let W_i be the customer i's waiting time. Another way to compute the total waiting time is to sum over each customer's waiting time, i.e.,

$$\mathbb{E}\left[\sum_{n=1}^{N} Z_n\right] = \mathbb{E}\left[\sum_{i=1}^{B} W_i\right], \tag{5.36}$$

where $B = \sum_{n=1}^{N} A_n$ is the total number of arrivals. Since the queue length is bounded by β, so

$$\mathbb{E}[W_i] \le \mathbb{E}[\zeta], \quad \text{where } \zeta = \min\left\{n' : \sum_{i=1}^{n'} D_i \ge \beta\right\}. \tag{5.37}$$

Combining (5.36) and (5.37), we have

$$\mathbb{E}\left[\sum_{n=1}^{N} Z_n\right] \le \mathbb{E}[\zeta] \sum_{i=1}^{B} 1 = O\left(\sqrt{N}\right),$$

since the total number of customers $B = \sum_{n=1}^{n} A_n = O\left(\sqrt{N}\right)$. □

Now, we bound the regret loss between our implemented policy and S-partial-SGD policy.

Proof. Proof of Proposition 5.3 By Theorem 5.2, $V(\delta, S)$ is Lipschitz in S with Lipschitz constant independent of δ. It follows that to show Proposition 5.3, it suffices to show

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(S^n - S^n_{j^n}\right)\right] = O(\sqrt{N}).$$

Let D^n denotes the first demand that occurs in the n^{th} epoch for $n = 1, 2, \ldots, N$. Consider

$$\begin{aligned}
S^{n+1} - S^{n+1}_{j^{n+1}} &= \max\left(S^{n+1}_{j^{n+1}}, x^{n+1}\right) - S^{n+1}_{j^{n+1}} = \left[x^{n+1} - S^{n+1}_{j^{n+1}}\right]^+ \le \left[S^n - D^n - S^{n+1}_{j^{n+1}}\right]^+ \\
&= \min_{j \in \mathcal{A}^{n+1}} \left[S^n - D^n - S^{n+1}_j\right]^+ = \min_{j \in \mathcal{A}^{n+1}} \left[S^n - D^n - \mathbf{Proj}_{[0,\beta]}\left(S^n_j - \eta_n \tilde{\nabla}^n_j\right)\right]^+ \\
&\le \min_{j \in \mathcal{A}^{n+1}} \left[S^n - D^n - \left(S^n_j - \eta_n \left|\tilde{\nabla}^n_j\right|\right)\right]^+ = \min_{j \in \mathcal{A}^{n+1}} \left[S^n - S^n_j - \left(D^n + \eta_n \left|\tilde{\nabla}^n_j\right|\right)\right]^+.
\end{aligned}$$

where the third equality is because $S^{n+1}_{j^{n+1}} = \max_{j \in \mathcal{A}^{n+1}} S^{n+1}_j$.
Case 1: If $j^n \in \mathcal{A}^{n+1}$, we have

$$S^{n+1} - S^{n+1}_{j^{n+1}} \le \left[S^n - S^n_{j^n} - \left(D^n + \eta_n \left|\tilde{\nabla}^n_{j^n}\right|\right)\right]^+.$$

By the definition of the stochastic gradient $\tilde{\nabla}^n_{j^n}$ in (5.10), we see that $\left|\tilde{\nabla}^n_{j^n}\right| \le \max(c, h) L^n$ where L^n is the length of the n^{th} epoch. That is

$$S^{n+1} - S^{n+1}_{j^{n+1}} \le [S^n - S^n + \eta_n \max(c, h) L^n - D^n]^+.$$

Case 2: If $j^n \notin \mathcal{A}^{n+1}$, we have

$$S^{n+1} - S^{n+1}_{j^{n+1}} \leq \Big[S^n - D^n - S^{n+1}_{j^{n+1}}\Big]^+ = \Big[S^n - S^n_{j^n} + (S^n_{j^n} - S^{n+1}_{j^{n+1}}) - D^n\Big]^+ \leq \Big[S^n - S^n_{j^n} + \beta - D^n\Big]^+,$$

where the last inequality is because both $S^n_{j^n}$ and $S^{n+1}_{j^{n+1}}$ are bounded in $[0,\beta]$.

Combing both cases, we have

$$S^{n+1} - S^{n+1}_{j^{n+1}} \leq \Big[S^n - S^n_{j^n} + \max(c,h)L^n + \beta\mathbb{1}\Big\{j^n \notin \mathcal{A}^{n+1}\Big\} - D^n\Big]^+.$$

Note $x_1 = 0$ implies $S^1 - S^1_{j^1} = 0$, and $S^n - S^n_{j^n} \in [0,\beta]$ for all $n = 1,\dots,N$. We can construct a queueing system that follows the structure of Lemma 6,

$$Z_{n+1} = \mathbf{Proj}_{[0,\beta]}\Big(Z_n + \eta_n \max(h,c)L^n + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n\Big),$$

and we have $Z_n \geq S^n - S^n_{j^n}$.

We note that the cycle length L^n and the first demand D^n in the n^{th} epoch are dependent. To simplify our analysis, we first decouple them as the follows. Let L^n_c be an *independent copy* of L^n. Then we claim that $\eta_n \max(h,c)L^n + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n$ is less than $\eta_n \max(h,c)(L^n_c + 1) + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n$ in distribution, i.e., for any $\alpha \in \mathbb{R}$, we have

$$\begin{aligned}&\mathbb{P}[\eta_n \max(h,c)L^n + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n \leq \alpha]\\ \leq&\mathbb{P}[\eta_n \max(h,c)(L^n_c + 1) + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n \leq \alpha]. \end{aligned} \tag{5.38}$$

This is because given any realization of $D^n = d^n$, the epoch length can be written as $L^n = 1 + \tau^n$, where τ^n is the hitting time for $\delta^n - d^n$. Apparently, τ^n is less than L^n_c in distribution, which implies that for any α,

$$\begin{aligned}&\mathbb{P}[\eta_n \max(h,c)L^n + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n \leq \alpha | D^n = d^n]\\ \leq&\mathbb{P}[\eta_n \max(h,c)(L^n_c + 1) + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n \leq \alpha | D^n = d^n]. \end{aligned} \tag{5.39}$$

Then (5.38) follows from (5.39) by unconditioning on D^n. Consider $\tilde{Z}_{n+1}$ defined by a queueing system

$$\tilde{Z}_{n+1} = \mathbf{Proj}_{[0,\beta]}\Big(\tilde{Z}_n + \eta_n \max(h,c)(L^n_c + 1) + \beta\mathbb{1}\{j^n \notin \mathcal{A}^{n+1}\} - D^n\Big),$$

and we obtain $\mathbb{E}[\sum_n \tilde{Z}_n] \geq \mathbb{E}[\sum_n Z_n]$.

Note that by Lemma 4, we have that L^n_c is (ν,b) sub-exponential, so by Corollary 5.9 of Azuma's inequality, $\sum_{n=1}^N \eta_n \max(h,c)(L^n_c + 1)$ is $(\max(h,c)\sqrt{\sum_{n=1}^N \eta_n^2 \nu},\ b)$-sub-exponential. For notational convenience, let $U = \sum_{n=1}^{N} \eta_n \max(h,c)(L_c^n + 1)$. Therefore,

$$\mathbb{P}[U \geq t + \mathbb{E}[U]] \leq \max\left\{ e^{-\frac{t^2}{2\nu'^2}}, e^{-\frac{t}{2b}} \right\},$$

where $\nu' = \max(h,c)\sqrt{\sum_{n=1}^{N} \eta_n^2 \nu}$. Equivalently, if we choose $t_0 = \max\left\{\nu'\sqrt{2\log N},\ 2b\log N\right\}$, then

$$\mathbb{P}[U \geq \mathbb{E}[U] + t_0] \leq \frac{1}{N}.$$

Therefore, define the event $A = \{U \geq t_0 + \mathbb{E}[U]\}$ and its complement A^c, then

$$\mathbb{E}\left[\sum_{n=1}^{N} \tilde{Z}_n\right] = \mathbb{P}[A]\mathbb{E}\left[\sum_{n=1}^{N} \tilde{Z}_n | A\right] + \mathbb{P}[A^c]\mathbb{E}\left[\sum_{n=1}^{N} \tilde{Z}_n | A^c\right] \leq \beta + \mathbb{E}\left[\sum_{n=1}^{N} \tilde{Z}_n | A^c\right]. \tag{5.40}$$

Note that conditioning on A^c, we have

$$\begin{aligned} U \leq t_0 + \mathbb{E}[U] &= \max\left\{\nu'\sqrt{2\log N}, 2b\log N\right\} + \mathbb{E}[U] \\ &= \max\left\{ \max(h,c)\sqrt{\sum_{n=1}^{N} \eta_n^2 \nu}\sqrt{2\log N}, 2b\log N \right\} + \mathbb{E}[U] = O\left(\sqrt{N}\right), \end{aligned}$$

where the last equality follows from plugging in the step size η_n and noticing $\mathbb{E}[L_c^n] \leq \mathbb{E}\left[\bar{L}\right]$.

On the other hand, $\sum_{n=1}^{N} \beta \mathbb{1}\{k^n \notin \mathcal{A}^{n+1}\} = O(\sqrt{N})$ because (a) the initial size of active set $\mathcal{A}^1 = \lfloor\sqrt{N}\rfloor$; (b) for any n, we have $\mathcal{A}^{n+1} \subseteq \mathcal{A}^n$. Thus, conditioning on A^c, $U + \beta\mathbb{1}\{k^n \notin \mathcal{A}^{n+1}\} = O(\sqrt{N})$. Note that D^n and L_c^n are independent, so conditioning on A^c, $\{D^n\}$ is still an i.i.d. sequence of random variables. By Lemma 6, we have $\mathbb{E}\left[\sum_{n=1}^{N} \tilde{Z}_n | A^c\right] = O(\sqrt{N})$. Plugging into (5.40) yields that $\mathbb{E}[\sum_{n=1}^{N} \tilde{Z}_n] = O(\sqrt{N})$, which then implies the desired result. □

This queueing system argument differs significantly from previous approaches (e.g., Huh and Rusmevichientong (2009); Shi et al. (2016)). To control the total amount of inventory deviation due to policy updates, they make stronger assumptions: Instead of assuming an upper bound on total inter-arrival times, i.e., $\sum_{n=1}^{N} A_n = O(\sqrt{N})$, they assume $\{A_n\}_{n=1}^{N}$ are independent to each other, and, for each $n = 1, 2 \ldots, N$, they bound the inter-arrival time as $A_n = O(1/\sqrt{n})$. In our setting, to bound the loss due to policy updates between epochs, we consider policy updates due to both running the stochastic gradient descent on S and running the bandit control on δ. Although the step size for SGD on S decays as $O(1/\sqrt{n})$, the bandit selection on δ has no such guarantees. Hence, we have to rely on bounding the total waiting time as in Lemma 6.

5.4.5 Proof of Proposition 5.4 – Bounding the SGD Loss

Via the crucial SAAP step, our algorithm keeps the number of SGD updates in a *synchronized* manner across all feasible policies within the active set. This enables us to bound the loss

$$\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n_{j^n}) - V^*_{j^n})\right].$$

We remark that if δ^n is kept constant for each epoch, then this term is the regret for applying the stochastic gradient algorithm to minimize some convex function, and, by the online convex optimization theory (Hazan et al. (2016)), it has regret $O(\sqrt{N})$. The subtle difference in our proof is to show that we can vary δ^n across epochs, and still maintain the same regret $O(\sqrt{N})$.

Proof. Proof of Proposition 5.4. Since $\mathbb{E}[L(\delta^n)] \geq 1$, we have

$$\mathbb{E}\left[\sum_{n=1}^{N}(V(\delta^n, S^n_{j^n}) - V^*_{j^n})\right] \leq \mathbb{E}\left[\sum_{n=1}^{N}(\mathbb{E}\left[G(\delta^n, S^n_{j^n}\right] - \min_{S}\mathbb{E}[G(\delta^n, S)])\right].$$

It suffices to show

$$\mathbb{E}\left[\sum_{n=1}^{N}(\mathbb{E}\left[G(\delta^n, S^n_{j^n})\right] - \min_{S}\mathbb{E}[G(\delta^n, S)])\right] = O(\sqrt{N}).$$

For notational convenience, we define the function $f^n_j(\cdot) := \mathbb{E}\left[G(\delta_j, \cdot)\right]$. Consider

$$\begin{aligned}2\mathbb{E}\left[\sum_{n=1}^{N}\left(f_{j^n}\left(S^n_{j^n}\right) - f_{j^n}\left(S^*_{j^n}\right)\right)\right] &\leq \mathbb{E}\left[\sum_{n=1}^{N}2\left\langle\nabla f_{j^n}\left(S^n_{j^n}\right), S^n_{j^n} - S^*_{j^n}\right\rangle\right]\\ &= \mathbb{E}\left[\sum_{n=1}^{N}2\left\langle\tilde{\nabla}^n_j, S^n_{j^n} - S^*_{j^n}\right\rangle\right]. \end{aligned} \tag{5.41}$$

By Pythagorean theorem,

$$\left\|S^{n+1}_j - S^*_j\right\|^2 = \left\|\mathbf{Proj}_{[\delta_j,\beta]}\left(S^n_j - \eta_n\tilde{\nabla}^n_j\right) - S^*_j\right\|^2 \leq \left\|S^n_j - \eta_n\tilde{\nabla}^n_j - S^*_j\right\|^2.$$

Hence, we have

$$\begin{aligned}\left\|S^{n+1}_j - S^*_j\right\|^2 &\leq \left\|S^n_j - S^*_j\right\|^2 - 2\eta_n\left\langle\tilde{\nabla}^n_j, S^n_j - S^*_j\right\rangle + (\eta_n)^2\left\|\tilde{\nabla}^n_j\right\|^2\\ 2\left\langle\tilde{\nabla}^n_j, S^n_j - S^*_j\right\rangle &\leq \frac{\left\|S^n_j - S^*_j\right\|^2 - \left\|S^{n+1}_j - S^*_j\right\|^2}{\eta_n} + \eta_n\left\|\tilde{\nabla}^n_j\right\|^2. \end{aligned} \tag{5.42}$$

Then by plugging (5.42) into (5.41) and setting $1/\eta_0 := 0$, we have

$$\begin{aligned}
2\mathbb{E}\left[\sum_{n=1}^{N}\left(f_{j^n}^{n}\left(S_{j^n}^{n}\right)-f_{j^n}^{n}\left(S_{j^n}^{*}\right)\right)\right] &\leq \mathbb{E}\left[\sum_{n=1}^{N}\frac{\left\|S_{j^n}^{n}-S_{j^n}^{*}\right\|^2-\left\|S_{j^n}^{n+1}-S_{j^n}^{*}\right\|^2}{\eta_n}+\eta_n\left\|\tilde{\nabla}_j^n\right\|^2\right] \\
&\leq \mathbb{E}\left[\sum_{n=1}^{N}\left\|S_{j^n}^{n}-S_{j^n}^{*}\right\|^2\left(\frac{1}{\eta_n}-\frac{1}{\eta_{n-1}}\right)\right]+\xi^2\sum_{n=1}^{N}\eta_n \\
&\leq \beta^2\sum_{n=1}^{N}\left(\frac{1}{\eta_n}-\frac{1}{\eta_{n-1}}\right)+\xi^2\sum_{n=1}^{N}\eta_n \\
&\leq 3\beta\xi\sqrt{N},
\end{aligned}$$

where the last inequality follows from that $\eta_n = \frac{\beta}{\xi\sqrt{n}}$ and $\sum_{n=1}^{N} 1/\sqrt{n} \leq 2\sqrt{N}$. □

5.5 Numerical Simulation

We conduct numerical experiments to study on the empirical performance of the proposed (δ, S) algorithm. The performance is evaluated by the percentage of increase in total cost of our algorithm π (over the planning horizon) compared with that of the clairvoyant optimal $\pi^* = (s^*, S^*)$ policy. That is, we measure the relative regret in terms of % as

$$r_T = \left(\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi}\right]-\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi^*}\right]\right)/\mathbb{E}\left[\sum_{t=1}^{T} C_t^{\pi^*}\right]\times 100\%.$$

Design of experiments. We first present the design of our numerical experiments, following Huh et al. (2009). The following parameters are common to all instances.

For the cost structure, we set the per-unit holding cost $h = 0.1$ and the per-unit ordering cost $c = 10$ and keep them unchanged. However, we vary the fixed cost $K = \{50, 100, 150\}$ and the per-unit lost-sales penalty cost $p = \{15, 25, 40\}$.

For the demand structure, we consider four commonly used demand distributions: (a) uniform, (b) gamma, (c) exponential, and (d) lognormal, with their respective parameters specified in Tables 5.1 and 5.2. The mean of all demand distributions is normalized to be 100 and the warehouse capacity is set to be $\beta = 1000$.

We run the (δ, S) learning algorithm over the planning horizons with number of periods $T \in \{100, 250, 500, 1000\}$. All systems start empty. For each testing instance, we generate 5000 sample paths of the random demand process, and use that to compute the average cost.

Numerical results. Under the four tested demand distributions, Tables 5.1 and 5.2 report the average computational performance of the (δ, S) learning algorithm. In addition, Tables 5.1 and 5.2 also give the sensitivity analysis with respect to the fixed cost K and the lost-sales penalty cost p, respectively.

Our key observations are summarized as follows. (a) The relative regret rates converge to zero consistently fast across all the tested demand distributions, which are well aligned with our analytical regret rate. For demand distributions that skew to the larger side, the algorithm converges faster. This is because the heavy-tail demand shortens the expected cycle length and thus increases the learning speed. (b) Our learning algorithm is robust with regard to the cost parameters K and p across all scenarios. Also, the results are consistent with our theoretical asymptotic regret rate.

distribution	fixed cost	optimal policy		optimal	relative regret r_T (%)			
	K	δ^*	S^*	average cost	125	250	500	1000
uniform	50	225.73	340.11	1015.76	6.38	5.67	4.97	4.92
	100	489.62	547.43	1030.49	4.66	4.25	3.61	3.52
	150	917.26	981.60	1036.74	5.17	3.88	2.95	2.86
gamma ($\alpha = 3$)	50	384.17	478.08	1015.05	5.08	5.05	4.98	4.65
	100	578.73	667.05	1031.77	4.56	3.79	3.64	3.53
	150	644.77	734.10	1038.75	4.50	3.74	3.65	3.50
gamma ($\alpha = 5$)	50	390.64	472.82	1017.20	5.30	5.20	4.85	4.65
	100	338.99	466.06	1029.54	5.80	4.37	3.80	3.22
	150	687.52	771.72	1041.12	4.11	3.37	3.28	3.06
gamma ($\alpha = 7$)	50	489.09	548.55	1018.23	4.84	4.00	3.99	3.72
	100	453.56	554.17	1030.30	5.11	3.87	3.28	2.76
	150	730.55	782.93	1043.23	3.85	3.01	2.83	2.81
exponential	50	599.53	648.39	1006.59	8.32	7.18	6.66	6.55
	100	703.03	751.85	1011.24	8.51	7.37	7.09	6.49
	150	756.28	851.91	1022.77	7.43	6.61	5.92	5.85
lognormal ($\sigma = 0.1$)	50	245.59	310.61	1025.98	4.25	3.07	2.55	2.32
	100	347.24	406.68	1040.84	3.14	2.01	1.75	1.65
	150	425.54	526.58	1049.70	3.38	2.05	1.43	1.21

Table 5.1: Performance of the (δ, S) algorithm with varying fixed costs K.

5.6 Conclusion

In this paper, we have proposed the first nonparametric learning algorithm for managing stochastic inventory systems with fixed costs under censored demand information, and showed that the cumulative regret is $O(\log T\sqrt{T})$, which is provably optimal up to a logarithmic factor. The algorithmic

distribution	lost-sale penalty p	optimal policy		optimal average cost	relative regret r_T (%)			
		δ^*	S^*		125	250	500	1000
uniform	15	249.60	384.00	1023.95	8.28	5.69	4.76	3.84
	25	625.31	787.87	1026.30	13.6	10.37	8.74	5.87
	40	687.06	889.28	1030.29	22.55	14.02	10.31	8.07
gamma ($\alpha = 3$)	15	413.59	504.41	1025.29	7.72	5.58	4.47	4.01
	25	552.07	756.41	1029.21	14.12	9.48	6.78	5.16
	40	556.34	764.71	1034.72	25.58	14.49	9.62	6.36
gamma ($\alpha = 5$)	15	287.51	458.2	1030.84	4.72	4.30	3.54	3.37
	25	481.31	625.7	1038.44	12.58	7.85	5.49	4.14
	40	451.96	632.94	1039.67	24.06	13.47	8.85	6.22
gamma ($\alpha = 7$)	15	315.98	420.32	1030.83	6.46	5.42	4.22	3.22
	25	493.99	641.01	1038.08	12.05	7.41	5.08	3.90
	40	570.39	709.08	1042.02	22.32	12.39	8.12	5.58
exponential	15	382.88	578.13	1027.26	9.72	7.37	6.53	6.03
	25	551.83	752.92	1038.18	16.83	11.16	7.85	7.21
	40	411.83	751.67	1026.86	22.31	14.62	10.76	8.04
lognormal ($\sigma = 1$)	15	438.18	522.72	1041.11	4.38	2.79	2.49	2.02
	25	439.5	542.91	1042.28	10.28	5.88	3.52	2.38
	40	410.96	519.90	1042.79	21.54	11.56	6.33	3.54

Table 5.2: Performance of the (δ, S) algorithm with varying lost-sale penalty costs p.

design and regret analysis involve several new and significant ideas that integrate the strength of stochastic gradient descent and bandit controls in a seamless fashion. The regret analysis consists of several bridging problems, and each pair of bridging problems requires a judiciously tuned hyper-parameters of the learning algorithm. We also develop several general technical results that are of independent interest beyond the context of this paper.

To close this paper, we point out two promising future research avenues. First, our framework allows for fixed costs, and there are many other important inventory systems that involve fixed cost. Second and more generally, our framework allows for multi-dimensional decision making and only requires *partial* convexity or concavity in the subset of decision variables and Lipschitz continuity in the remaining set. There are many interesting applications arising in the context of operations management, e.g., dual-sourcing inventory systems, joint inventory and pricing control, joint inventory and vehicle routing, joint scheduling and medical decision making.

Appendix.

5.7 A Summary of Major Notation

Table 5.3 summarizes the major mathematical notation used in the manuscript.

t	index of periods
i	index of cycles
n	index of epochs
h	unit holding cost
p	unit lost-sales penalty cost
K	fixed setup cost
c	unit ordering cost
D_t, d_t	demand and realized demand in periods t
x_t	beginning inventory in period t
q_t	ordering quantity in period t
y_t	after ordering inventory in periods t, i.e., $y_t = x_t + q_t$
C_t	cost in period t
$\tilde{C}_t$	pseudo cost in period t
(δ, S) policy	δ: inventory gap, S: order-up-to level
$\{\delta_j\}_{j=1}^{J}$	discrete inventory gaps
j	index for discrete inventory gaps
(δ^n, S^n)	implemented policy in the n^{th} epoch
$L(\delta, S) = L(\delta)$	random cycle length for (δ, S) policy
$L(\delta, S, \bar{\delta})$	random epoch length with parameters $(\delta, S, \bar{\delta})$
$H(\delta, S)$	random cycle cost for (δ, S) policy
$G(\delta, S)$	random cycle pseudo cost for (δ, S) policy
$\tilde{G}(\delta, S, \bar{\delta})$	random epoch pseudo cost with parameters $(\delta, S, \bar{\delta})$
$V(\delta, S)$	long-run average pseudo cost for (δ, S) policy
$V^*(\delta) = \min_S V(\delta, S)$	long-run average pseudo cost for δ with optimal S

$V^* = \min_{\delta,S} V(\delta,S)$	optimal long-run average pseudo cost
$\hat{V}^n(\delta)$	n-step approximation for $V^*(\delta)$
$\hat{G}^n(\delta)$	n-step cumulative cycle pseudo cost along SGD path
$\hat{L}^n(\delta)$	n-step cumulative cycle length
$L_j(S) = L(\delta_j, S)$ $G_j(S) = G(\delta_j, S)$ $V_j(S) = V(\delta_j, S)$ $V_j^* = V(\delta_j, S_j^*)$ $\hat{V}_j^n = \hat{V}^n\left(\delta_j\right)$ $\hat{G}_j^n = \hat{G}^n\left(\delta_j\right)$ $\hat{L}_j^n = \hat{L}^n\left(\delta_j\right)$	using sub-script j to represent δ_j argument
$\tilde{\nabla}_j^n$	stochastic gradient for inventory gap δ_j in the n^{th} epoch
$\bar{\delta}^n$	largest inventory gap in the n^{th} epoch
j^n	index of policy with largest S in the n^{th} epoch
η_n	SGD step size in the n^{th} epoch
Δ^n	confidence size in the n^{th} epoch
$\mathcal{D}^n$	list of censored demand data in the n^{th} epoch
$\mathcal{A}^n$	policy active set in the n^{th} epoch
x^n	inventory level at the beginning of the n^{th} epoch
β	max inventory capacity
γ	max per period cost
ξ^2	bound second moment for stochastic gradient
(ν, b)	universal sub-exponential parameters for cycle length, cost, and stochastic gradient for all (δ, S) policy
θ	constant for controlling estimation error from $\hat{V}_j^n$ to V_j^*
$\bar{L} = L(\beta)$	cycle length with maximum inventory gap

Table 5.3: Summary of Major Notation

5.8 Technical Proofs for Results in Section 5.3

Proof. Proof of Lemma 2. By (5.7) and the storage capacity β, it is clear that $|\tilde{C}_t| \leq K + (h + c + p)\beta$. □

Proof. Proof of Lemma 3. Let f be the pdf of D_1. Consider

$$\begin{aligned} \mathbb{E}[L(\delta)] &= \sum_{t=1}^{\infty} \mathbb{P}[L(\delta) \geq t] = \sum_{t=1}^{\infty} \mathbb{P}[W_{t-1} < \delta] \\ &= 1 + \sum_{t=1}^{\infty} \int_0^{\delta} f_{W_t}(x)dx = 1 + \int_0^{\delta} \sum_{t=1}^{\infty} f_{W_t}(x)dx = 1 + \int_0^{\delta} \sum_{t=1}^{\infty} f^{*t}(x)dx, \end{aligned} \tag{5.43}$$

where the pdf of W_t is denoted by f_{W_t} and $f_{W_t} = f * f * \cdots * f = f^{*t}$ (i.e., the convolution of f's), and the interchange between summation and integration in the fourth equality is backed by the monotone convergence theorem.

From (5.43), we see that $\frac{d}{d\delta}\mathbb{E}[L(\delta)] = \sum_{t=1}^{\infty} f^{*t}(\delta)$. Hence, to show $\mathbb{E}[L(\delta)]$ is Lipschitz with Lipschitz constant 6ρ, it suffices to show its derivative $\sum_{t=1}^{\infty} f^{*t}(x) \leq 6\rho$ for all $x \geq 0$.

Let m be the median of D_1, and we partition $f = f^l + f^r$, where

$$f^l(x) = \begin{cases} f(x), \text{ if } x \leq m, \\ 0, \text{ otherwise;} \end{cases} \qquad f^r(x) = \begin{cases} f(x), \text{ if } x > m, \\ 0, \text{ otherwise.} \end{cases}$$

Define $F_n = \sum_{t=1}^{n} f^{*t}$ and $Z_n = \sup_x F_n(x)$. We want to show that $Z_n \leq 6\rho$ for all n. We have

$$F_n \leq F_{n+1} = f + F_n * f = f + F_n * f^l + F_n * f^r. \tag{5.44}$$

By using the definition of Z_n and the fact that m is the median of D, we have

$$F_n * f^r(x) = \int_{-\infty}^{\infty} F_n(y) f^r(x-y)dx \leq Z_n \int_m^{\infty} f^r(x-y)dy = Z_n/2. \tag{5.45}$$

for all $x \geq 0$. Plugging (5.45) into (5.44), we have that for all $n \geq 1$,

$$F_n \leq \rho + Z_n/2 + \sup_x F_n * f^l(x) \leq \rho + Z_n/2 + \sup_x (f^l + F_\infty * f^l)(x),$$

Since Z_n is the supreme over F_n, then we have

$$Z_n \leq \rho + Z_n/2 + \sup_x (f^l + F_\infty * f^l)(x), \tag{5.46}$$

which implies that

$$Z_n \le 2\left(\rho + \sup_x (f^l + F_\infty * f^l)(x)\right). \tag{5.47}$$

We shall show $f^l + F_\infty * f^l < 2\rho$ uniformly. To prove this, for any fixed small positive constant ϵ, we define τ to be the hitting time of target interval $[\delta, \delta+\epsilon]$, i.e., $\tau = \min\{t : W_t \in [\delta, \delta+\epsilon]\}$. Consider the probability of event that we hit $[\delta, \delta+\epsilon]$ from a distance less than m. We have

$$\begin{aligned}\mathbb{P}[D_\tau < m] &= \sum_{i=1}^{\infty} \mathbb{P}[\tau = i, D_i < m] = \sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} < \delta, W_i \in [\delta, \delta+\epsilon], D_i < m] \\ &\ge \sum_{i=1}^{\infty} \mathbb{P}[W_i \in [\delta, \delta+\epsilon], D_i < m] - \sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} \in [\delta, \delta+\epsilon], W_i \in [\delta, \delta+\epsilon]]. \end{aligned} \tag{5.48}$$

We first focus on the first term on the RHS of (5.48).

$$\sum_{i=1}^{\infty} \mathbb{P}[W_i \in [\delta, \delta+\epsilon], D_i < m] = \sum_{i=1}^{\infty} \int_{\delta}^{\delta+\epsilon} \left(f^{*(i-1)} * f^l\right) = \int_{\delta}^{\delta+\epsilon} \sum_{i=1}^{\infty} \left(f^{*(i-1)} * f^l\right) = \int_{\delta}^{\delta+\epsilon} \left(f^l + F_\infty * f^l\right),$$

where the interchange between summation and integration is backed by the monotone convergence theorem. Next, we focus on the second term on the RHS of (5.48).

$$\begin{aligned}\sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} \in [\delta, \delta+\epsilon], W_i \in [\delta, \delta+\epsilon]] &= \sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} \in [\delta, \delta+\epsilon]] \mathbb{P}[W_i \in [\delta, \delta+\epsilon] | W_{i-1} \in [\delta, \delta+\epsilon]] \\ &\le \sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} \in [\delta, \delta+\epsilon]] \mathbb{P}[D_i \le \epsilon | W_{i-1} \in [\delta, \delta+\epsilon]] \\ &\le \sum_{i=1}^{\infty} \mathbb{P}[W_{i-1} \in [\delta, \delta+\epsilon]] \rho\epsilon \\ &= \rho\epsilon \sum_{i=1}^{\infty} \int_{\delta}^{\delta+\epsilon} f_{W_{i-1}} = \rho\epsilon \int_{\delta}^{\delta+\epsilon} \sum_{i=1}^{\infty} f_{W_{i-1}} = \rho\epsilon \int_{\delta}^{\delta+\epsilon} F_\infty, \end{aligned}$$

where, again, the interchange between summation and integration is backed by the monotone convergence theorem. Putting the above two terms in (5.48), we have

$$\mathbb{P}[D_i < m] \ge \int_{\delta}^{\delta+\epsilon} \left(f^l + F_\infty * f^l\right) - \rho\epsilon \int_{\delta}^{\delta+\epsilon} F_\infty. \tag{5.49}$$

On the other hand, we define $\tau' = \min\{t : W_t > \delta - m\}$, which is the first time that the random walk

W_t crosses $\delta - m$. Consider

$$\mathbb{P}[D_\tau < m|W_{\tau'}] = \sum_{i=1}^{\infty} \mathbb{P}[\tau = \tau' + i, D_\tau < m|W_{\tau'}] \leq \sum_{i=1}^{\infty} \frac{1}{2^{i-1}} \epsilon\rho = 2\epsilon\rho, \tag{5.50}$$

where the last inequality holds due to the following argument. If $\tau = \tau' + i$ and $D_\tau < m$, we must have $D_{\tau'+i'} < m$ for $i' = 1, \ldots, i-1$ and $D_{\tau+i}$ must hit a target interval of size ϵ, which has probability no more than $\frac{1}{2^{i-1}}\rho\epsilon$ (as m is the median, and ρ is an upper bound on densities). Hence,

$$\mathbb{P}[D_\tau < m] = \mathbb{E}[\mathbb{P}[D_\tau < m|W_{\tau'}]] \leq 2\epsilon\rho. \tag{5.51}$$

Combining (5.49) and (5.51), we have

$$\int_{\delta}^{\delta+\epsilon} \left(f^l + F_\infty * f^l\right) - \rho\epsilon \int_{\delta}^{\delta+\epsilon} F_\infty \leq 2\epsilon\rho.$$

Dividing both sides by ϵ, and taking $\epsilon \to 0$, we have

$$f^l + F_\infty * f^l \leq 2\rho. \tag{5.52}$$

Plugging (5.52) into (5.47), we have $Z_n \leq 6\rho$ for all n, which yields the desired result. □

Proof. Proof of Theorem 5.2. Recall that $V(\delta, S) := \mathbb{E}[G(\delta, S)]/\mathbb{E}[L(\delta, S)]$, and $L(\delta, S)$ counts the number of periods until the cumulative demand in the cycle exceeds δ, which is independent of S. Therefore, to show $V(\delta, S)$ is convex in S, it suffices to show that $\mathbb{E}[G(\delta, S)]$ is convex in S. Consider a cycle with demand samples $d_1, \ldots, d_L$. The cycle cost can be written as

$$\begin{aligned} G(\delta, S) &= \sum_{t=1}^{L} \left[h(x_t + q_t - d_t)^+ - p\min(x_t + q_t, d_t)\right] + K + c(S - x_{L+1}) \\ &= \begin{cases} K + \sum_{t=1}^{L}(hx_{t+1} - pd_t) + c(S - x_{L+1}), & \text{if } x_{L+1} > 0 \text{ and } L \geq 1, \\ K + \sum_{t=1}^{L-1}(hx_{t+1} - pd_t) - px_L + cS, & \text{if } x_{L+1} = 0 \text{ and } L > 1, \\ K - pS + cS, \quad \text{if } x_{L+1} = 0 \text{ and } L = 1, \end{cases} \end{aligned} \tag{5.53}$$

where $x_{t+1} = \max(S - d_1 - \ldots - d_t, 0)$ for $t = 1, \ldots, L$. Then taking derivative with respect to S,

$$\nabla_S G(\delta, S) = \begin{cases} hL, & \text{if } x_{L+1} > 0, \\ h(L-1) - p + c, & \text{if } x_{L+1} = 0, \end{cases} \tag{5.54}$$

which gives an unbiased stochastic (S-partial) gradient of $\mathbb{E}[G(\delta, S)]$. Because $|\nabla_S G(\delta, S)|$ is clearly bounded almost surely, we also have

$$\begin{aligned}\nabla_S \mathbb{E}[G(\delta,S)] =&\mathbb{E}[hL\mathbb{1}(x_{L+1}>0)+(h(L-1)-p+c)\mathbb{1}(x_{L+1}=0)]\\ =&\mathbb{E}[hL+(-h-p+c)\mathbb{1}(x_{L+1}=0)]\\ =&h\mathbb{E}[L]-(h+p-c)\mathbb{P}(x_{L+1}=0).\end{aligned}$$

Since $\mathbb{E}[L]$ is independent of S, $p>c$ in the lost-sales model, and $\mathbb{P}(x_{L+1}=0)$ is decreasing in S, we conclude that $\nabla_S \mathbb{E}[G(\delta,S)]$ is increasing in S, which implies that $\mathbb{E}[G(\delta,S)]$ is convex in S. Also, by (5.54) and $p>c$, we have

$$|\mathbb{E}[\nabla_S G(\delta,S))]| \le h\mathbb{E}\left[\bar{L}\right]+p, \tag{5.55}$$

where $\bar{L}$ denotes the cycle length associated with the maximum $\delta=\beta$ (which is independent of S). Then, since $\mathbb{E}[L(\delta)]\ge 1$, we have $|\nabla_S V(\delta,S)|\le h\mathbb{E}\left[\bar{L}\right]+p+c$, and therefore $V(\delta,S)$ is Lipschitz in S, and the Lipschitz constant is independent of S.

Now, we show that $V^*(\delta)=\min_{S\in[\delta,\beta]}V(\delta,S)$ is Lipschitz in δ. For any two inventory gaps δ_1 and δ_2, without loss of generality, we assume that $V^*(\delta_1)\ge V^*(\delta_2)$. Define the corresponding minimizers $S_1=\arg\min_{S\in[\delta_1,\beta]}V(\delta_1,S)$ and $S_2=\arg\min_{S\in[\delta_2,\beta]}V(\delta_2,S)$. Then, we have

$$V^*(\delta_1)-V^*(\delta_2)=V(\delta_1,S_1)-V(\delta_2,S_2)\le V(\delta_1,\max\{S_2,\delta_1\})-V(\delta_2,S_2).$$

For better readability, we slight abuse the notation to define

$$G_1=\mathbb{E}[G(\delta_1,\max\{\delta_1,S_2\})],\ L_1=\mathbb{E}[L(\delta_1,\max\{S_2,\delta_1\})],\ G_2=\mathbb{E}[G(\delta_2,S_2)],\ L_2=\mathbb{E}[L(\delta_2,S_2)].$$

Then we have

$$\begin{aligned}&V(\delta_1,\max\{\delta_1,S_2\})-V(\delta_2,S_2)\\ =&G_1/L_1-G_2/L_2=G_1/L_1-G_2/L_1+G_2/L_1-G_2/L_2=\frac{1}{L_1}(G_1-G_2)+\frac{G_2}{L_1L_2}(L_1-L_2).\end{aligned} \tag{5.56}$$

We first analyze the term G_1-G_2. If $S_2\ge\delta_1$, by Lemma 2, since G_1 and G_2 share the same starting point S_2, we have $G_1-G_2\le\gamma|L_1-L_2|$. On the other hand, if $S_2<\delta_1$,

$$\begin{aligned}G_1-G_2&=\mathbb{E}[G(\delta_1,\delta_1)]-\mathbb{E}[G(\delta_2,S_2)]\\ &=\mathbb{E}[G(\delta_1,\delta_1)]-\mathbb{E}[G(\delta_2,\delta_1)]+\mathbb{E}[G(\delta_2,\delta_1)]-\mathbb{E}[G(\delta_2,S_2)]\\ &\le\gamma|L_1-L_2|+|\delta_1-S_2|(h\mathbb{E}\left[\bar{L}\right]+p)\end{aligned}$$

$$\leq \gamma |L_1 - L_2| + |\delta_1 - \delta_2| (h\mathbb{E}\left[\bar{L}\right] + p), \tag{5.57}$$

where the first inequality is due to (5.55), and the last inequality is due to $\delta_2 \leq S_2 \leq \delta_1$. So (5.57) holds true for both cases. Plugging (5.57) into (5.56), we have

$$\begin{aligned} V(\delta_1, S_1) - V(\delta_2, S_2) &\leq \frac{1}{L_1}\left(\gamma |L_1 - L_2| + |\delta_1 - \delta_2| (h\mathbb{E}\left[\bar{L}\right] + p)\right) + \left|\frac{G_2}{L_1 L_2}(L_1 - L_2)\right| \\ &\leq (h\mathbb{E}\left[\bar{L}\right] + p) |\delta_1 - \delta_2| + 2\gamma |L_1 - L_2|. \end{aligned}$$

By Lemma 3, we have $|L_1 - L_2| \leq 6\rho |\delta_1 - \delta_2|$. Therefore, we conclude that

$$V(\delta_1, S_1) - V(\delta_2, S_2) \leq \left(h\mathbb{E}\left[\bar{L}\right] + p + 12\gamma\rho\right) |\delta_1 - \delta_2|,$$

which shows that $V^*(\delta)$ is Lipschitz in δ. □

5.9 Known Results on Sub-Exponential Random Variables

The standard results in this section are stated without proofs, and we refer interested readers to Wainwright (2019) for their detailed arguments.

Definition 6 (Martingale Difference Sequence). *A martingale difference sequence is an adapted sequence $\{X_k, \mathcal{F}_k\}_{k=0}^{\infty}$ such that for all $k \geq 1$,*

$$\mathbb{E}[|X_k|] < \infty, \text{ and } \mathbb{E}[X_k | \mathcal{F}_k] = 0.$$

Definition 7 (Sub-Exponential Random Variable). *A random variable X with mean $\mu = \mathbb{E}[X]$ is called sub-exponential if there are non-negative parameters (ν, b) such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < 1/b.$$

Definition 8 (Sub-Exponential Random Vector). *A random vector $X \in \mathbb{R}^d$ is said to be sub-exponential with parameters (ν, b) if for any unit vector $u \in \mathbb{R}^d$, the random variable $u^T(X - \mathbb{E}[X])$ is (ν, b) sub-exponential.*

Theorem 5.6 (Sub-Exponential Tail Bound). *Suppose X is sub-exponential with parameters (ν, b) and mean $\mu = \mathbb{E}[X]$. Then we have*

$$\mathbb{P}[X \geq \mu + t] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ e^{-\frac{t}{2b}} & \text{if } t > \frac{\nu^2}{b}, \end{cases}$$

and

$$\mathbb{P}[|X-\mu| \geq t] \leq \begin{cases} 2e^{-\frac{t^2}{2\nu^2}} & \textit{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ 2e^{-\frac{t}{2b}} & \textit{if } t > \frac{\nu^2}{b}. \end{cases}$$

Theorem 5.7 (Equivalent Characterization of Sub-Exponentials). *For a zero-mean random variable X, the following statements are equivalent:*

1. *There are non-negative numbers* (ν, b) *such that*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \textit{for all } |\lambda| < 1/b.$$

2. *There are constants* $c_1, c_2 > 0$ *such that*

$$\mathbb{P}[|X| > t] \leq c_1 e^{-c_2 t} \quad \textit{for all } t > 0.$$

Theorem 5.8 (Azuma's Inequality). *Suppose* $\{(X_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ *is a martingale difference sequence, and for any* $|\lambda| < 1/b_k$,

$$\mathbb{E}\left[e^{\lambda X_k} | \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2} \textit{ almost surely.}$$

Then $\sum_{k=1}^{n} a_k X_k$ *is sub-exponential with parameters* $(\sqrt{\sum_{k=1}^{n} a_k^2 \nu_k^2}, \max_{k\in[n]} b_k)$. *Consequently, for all* $t \geq 0$,

$$\mathbb{P}\left[\sum_{k=1}^{n} a_k X_k \geq t\right] \leq \begin{cases} e^{-\frac{t^2}{2\sum_{k=1}^{n} a_k^2 \nu_k^2}} & \textit{if} \quad 0 \leq t \leq \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \\ e^{-\frac{t}{2\max_{k\in[n]} b_k}} & \textit{if} \quad t > \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \end{cases}$$

and

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} a_k X_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^{n} a_k^2 \nu_k^2}} & \textit{if} \quad 0 \leq t \leq \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \\ 2e^{-\frac{t}{2\max_{k\in[n]} b_k}} & \textit{if} \quad t > \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}. \end{cases}$$

Theorem 5.9 (Corollary of Azuma's Inequality). *Suppose* $\{X_k\}_{k=1}^{n}$ *are independent centered random variables, and* X_k *is sub-exponential with parameters* (ν_k, b_k). *Then* $\sum_{k=1}^{n} a_k X_k$ *is sub-exponential with parameters* $(\sqrt{\sum_{k=1}^{n} a_k^2 \nu_k^2}, \max_{k\in[n]} b_k)$. *Consequently, for all* $t \geq 0$,

$$\mathbb{P}\left[\sum_{k=1}^{n} a_k X_k \geq t\right] \leq \begin{cases} e^{-\frac{t^2}{2\sum_{k=1}^{n} a_k^2 \nu_k^2}} & \textit{if} \quad 0 \leq t \leq \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \\ e^{-\frac{t}{2\max_{k\in[n]} b_k}} & \textit{if} \quad t > \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \end{cases}$$

and

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} a_k X_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^{n} a_k^2 \nu_k^2}} & \textit{if} \quad 0 \leq t \leq \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}, \\ 2e^{-\frac{t}{2\max_{k\in[n]} b_k}} & \textit{if} \quad t > \frac{\sum_{k=1}^{n} a_k^2 \nu_k^2}{\max_{k\in[n]} b_k}. \end{cases}$$

CHAPTER 6

Conclusion and Future Work

This dissertation focuses on the design and analysis of data-driven algorithms for stochastic inventory and supply chain systems. To design learning algorithms with theoretical performance guarantees, we focus on learning simple structured policies. We decouple system dynamics with *cycling trick*, and design variations of online learning algorithms. More concretely, we extend the confidence bound approaches from various perspectives: We develop algorithms of confidence-bound type which estimate expected cycle average cost in Chapter 1, optimal value for convex cost function in Chapters 2 and 5, and long-run limiting cost in Markovian setting in Chapter 3. Moreover, in Chapters 2 and 5, we handle the censored demand issue by developing a shrinking active set algorithm which leverages the one-side zeroth ordering information. In Chapter 4, we customize the mirror decent algorithm to improve the high dimensional inventory decisions for multi-production supply chain system. Besides these learning algorithms, we also develop a new queuing system theory which helps us design data-driven algorithms adaptive to physical consultations.

To close this Ph.D. thesis, we would like to point out several promising and important future research avenues. First, one may employ or further innovate the methods developed here to study more complex systems, e.g., the dual-sourcing inventory control problem, the joint inventory-location problem, the inventory routing problem, and the joint pricing and inventory control problem. Second, instead of regret framework, one may re-examine the problems using in the best-arm identification setting, i.e., figuring out the clairvoyant optimal policies as fast as possible (without being overly concerned with cumulative regret). Third, one may integrate the theory of Markov chain mixing time and the theory of high-dimensional statistical learning into our current framework to tackle more difficult but important problems.

BIBLIOGRAPHY

Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *JMLR Workshop and Conference Proceedings*, volume 40. Microtome Publishing.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Azoury, K. S. (1985). Bayes solution to dynamic inventory models under unknown demand distribution. *Management science*, 31(9):1150–1160.

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

Beyer, D., Sethi, S. P., and Sridhar, R. (2001). Stochastic multiproduct inventory models with limited storage. *Journal of Optimization Theory and Applications*, 111(3):553–588.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695.

Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. (2009). Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 201–208.

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

Caliskan-Demirag, O., Chen, Y., and Yang, Y. (2012). Ordering policies for periodic-review inventory systems with quantity-dependent fixed costs. *Operations Research*, 60(4):785–796.

Chao, X. and Zipkin, P. H. (2008). Optimal policy for a periodic-review inventory system under a supply capacity contract. *Operations Research*, 56(1):59–68.

Chen, B., Chao, X., and Ahn, H.-S. (2018a). Coordinating pricing and inventory replenishment with nonparametric demand learning. *To appear in Operations Research.*

Chen, B., Chao, X., and Shi, C. (2018b). Nonparametric algorithms for joint pricing and inventory control with lost-sales and censored demand. *Working Paper. University of Michigan, Ann Arbor, MI, USA*.

Chen, L. and Plambeck, E. L. (2008). Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing & Service Operations Management*, 10(2):236–256.

Chen, S. (2004). The infinite horizon periodic review problem with setup costs and capacity constraints: A partial characterization of the optimal policy. *Operational Research*, 52(3):409–421.

Chen, S. and Lambrecht, M. (1996). X-Y band and modified (s, S) policy. *Operations Research*, 44:1013–1019.

Chen, W., Shi, C., and Duenyas, I. (2018c). Optimal learning algorithms for stochastic inventory systems with random capacities. *Available at SSRN 3287560*.

Chen, W., Shi, C., and Duenyas, I. (2018d). Optimal learning algorithms for stochastic inventory systems with random capacities. *Working Paper. University of Michigan, Ann Arbor, MI*.

Chen, X. and Simchi-Levi, D. (2004a). Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. *Operations Research*, 52(6):887–896.

Chen, X. and Simchi-Levi, D. (2004b). Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The infinite horizon case. *Mathematics of operations Research*, 29(3):698–723.

Chen, Y., Ray, S., and Song, Y. (2006). Optimal pricing and inventory control policy in periodic-review systems with fixed ordering cost and lost sales. *Naval Research Logistics (NRL)*, 53(2):117–136.

Cheung, M., Elmachtoub, A. N., Levi, R., and Shmoys, D. B. (2016). The submodular joint replenishment problem. *Mathematical Programming*, 158(1-2):207–233.

Chu, L. Y., Shanthikumar, J. G., and Shen, Z.-J. M. (2008). Solving operational statistics via a bayesian analysis. *Operations Research Letters*, 36(1):110–116.

Federgruen, A. and Zipkin, P. (1984). An efficient algorithm for computing optimal (s, S) policies. *Operations research*, 32(6):1268–1285.

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

Gallego, G. and Özer, Ö. (2001). Integrating replenishment decisions with advance demand information. *Management science*, 47(10):1344–1360.

Gallego, G. and Scheller-Wolf, A. (2000). Capacitated inventory problems with fixed order costs: Some optimal policy structure. *European Journal of Operational Research*, 126:603–613.

Gavirneni, S. (2001). An efficient heuristic for inventory control when the customer is using a (s, S) policy. *Operations Research Letters*, 28(4):187–192.

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

Guan, Y. and Miller, A. J. (2008). Polynomial-time algorithms for stochastic uncapacitated lot-sizing problems. *Operations Research*, 56(5):1172–1183.

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

Hu, P., Lu, Y., and Song, M. (2018). Joint pricing and inventory control with fixed and convex/concave variable production costs. *To appear in Production and Operations Management*.

Huang, K. and KüçüKyavuz, S. (2008). On stochastic lot-sizing problems with random lead times. *Operations Research Letters*, 36(3):303–308.

Huh, W. T. and Janakiraman, G. (2008). (s, S) optimality in joint inventory-pricing control: An alternate approach. *Operations Research*, 56(3):783–790.

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

Huh, W. T., Levi, R., Rusmevichientong, P., and Orlin, J. B. (2011). Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4):929–941.

Huh, W. T. and Rusmevichientong, P. (2009). A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123.

Iglehart, D. L. (1963). Optimality of (s, S) policies in the infinite horizon dynamic inventory problem. *Management science*, 9(2):259–267.

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

Ignall, E. and Veinott Jr, A. F. (1969). Optimality of myopic inventory policies for several substitute products. *Management Science*, 15(5):284–304.

Khouja, M. and Goyal, S. (2008). A review of the joint replenishment problem literature: 1989–2005. *European Journal of Operational Research*, 186(1):1–16.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

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

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lee, Y. T. and Vempala, S. (2019). *Techniques in Optimization and Sampling*. Working book: https://www.dropbox.com/s/10zwtzolc1qhpsq/main.pdf.

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

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

Levi, R. and Shi, C. (2013). Approximation algorithms for the stochastic lot-sizing problem with order lead times. *Operations Research*, 61(3):593–602.

Liyanage, L. H. and Shanthikumar, J. G. (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341–348.

Lu, X., Song, J.-S., and Zhu, K. (2005). On "the censored newsvendor and the optimal acquisition of information". *Operations Research*, 53(6):1024–1026.

Lu, X., Song, J.-S., and Zhu, K. (2008). Analysis of perishable-inventory systems with censored demand data. *Operations Research*, 56(4):1034–1038.

Murray, G. R. and Silver, E. A. (1966). A bayesian analysis of the style goods inventory problem. *Management Science*, 12(11):785–797.

Nagarajan, V. and Shi, C. (2016). Approximation algorithms for inventory problems with submodular or routing costs. *Mathematical Programming*, 160(1-2):225–244.

Özer, Ö. and Wei, W. (2004). Inventory control with limited capacity and advance demand information. *Operations Research*, 52(6):988–1000.

Pang, Z., Chen, F. Y., and Feng, Y. (2012). A note on the structure of joint inventory-pricing control with leadtimes. *Operations Research*, 60(3):581–587.

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

Ross, S. M. (1996). *Stochastic Processes. 2nd Edition.* John Wiley & Sons, New York, NY.

Scarf, H. (1959). Bayes solutions of the statistical inventory problem. *The annals of mathematical statistics*, 30(2):490–508.

Scarf, H. (1960). *The optimality of (S, s) policies in the dynamic inventory problem.* Mathematical Methods in the Social Sciences. Stanford University Press, Stanford, CA.

Scheller-Wolf, A., Veeraraghavan, S., and van Houtum, G.-J. (2007). Effective dual sourcing with a single index policy. In *Working paper, Carnegie Mellon University, Pittsburgh, PA*. Citeseer.

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

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

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

Shi, C., Zhang, H., Chao, X., and Levi, R. (2014). Approximation algorithms for capacitated stochastic inventory systems with setup costs. *Naval Research Logistics (NRL)*, 61(4):304–319.

Simchi-Levi, D., Chen, X., and Bramel, J. (2014). *The Logic of Logistics: Theory, Algorithms, and Applications for Logistics and Supply Chain Management*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY.

Veinott Jr, A. F. (1966). The status of mathematical inventory theory. *Management Science*, 12(11):745–777.

Veinott Jr, A. F. and Wagner, H. M. (1965). Computing optimal (s, S) inventory policies. *Management Science*, 11(5):525–552.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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

Zhang, H., Chao, X., and Shi, C. (2019). Closing the gap: a learning algorithm for lost-sales inventory systems with lead times. *To appear in Management Science*.

Zheng, Y.-S. (1991). A simple proof for optimality of (s, S) policies in infinite-horizon inventory systems. *Journal of Applied Probability*, 28(4):802–810.

Zheng, Y.-S. and Federgruen, A. (1991). Finding optimal (s, S) policies is about as simple as evaluating a single policy. *Operations research*, 39(4):654–665.

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

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