No-Regret Learning in Two-Echelon Supply Chain with Unknown Demand Distribution

Mengxiao Zhang[†] **Shi Chen**[‡] **Haipeng Luo**[†] **Yingfei Wang**[‡]
University of Southern California[†]
University of Washington[‡]

Abstract

Supply chain management (SCM) has been recognized as an important discipline with applications to many industries, where the two-echelon stochastic inventory model, involving one downstream retailer and one upstream supplier, plays a fundamental role for developing firms' SCM strategies. In this work, we aim at designing online learning algorithms for this problem with an unknown demand distribution, which brings distinct features as compared to classic online optimization problems. Specifically, we consider the two-echelon supply chain model introduced in (Cachon and Zipkin, 1999) under two different settings: the *centralized* setting, where a planner decides both agents' strategy simultaneously, and the *decentralized* setting, where two agents decide their strategy independently and selfishly. We design algorithms that achieve favorable guarantees for both regret and convergence to the optimal inventory decision in both settings, and additionally for individual regret in the decentralized setting. Our algorithms are based on Online Gradient Descent and Online Newton Step, together with several new ingredients specifically designed for our problem. We also implement our algorithms and show their empirical effectiveness.

1 Introduction

A supply chain is two or more parties linked by a flow of goods, information, and funds, before a product can be finally delivered to outside customers. When multiple decision makers are involved, behavior that is locally rational can be inefficient from a global perspective. Supply chain management (SCM) research then focuses on methods for improving system efficiencies, so as to "efficiently integrate suppliers, manufacturers, warehouses, and stores $\cdots$ in order to minimize system-wide costs while satisfying service level requirements" (Simchi-Levi et al., 1999). In the vast body of SCM literature, the mathematical model of a two-echelon stochastic inventory system with a known demand distribution plays a fundamental role for analyzing firms' SCM strategies and has been well studied over the past decades (Clark and Scarf, 1960; Federgruen and Zipkin, 1984; Chen and Zheng, 1994; Cachon and Zipkin, 1999).

In the classic two-echelon stochastic inventory planning problem, two agents, Agent 1 (the retailer, referred to as *he*) and Agent 2 (the supplier, referred to as *she*), will go through a process of T rounds. Following the sequence of events in the SCM literature (Cachon and Zipkin, 1999), Agent 1 first observes an external demand $d_t \sim \mathcal{D}$ and utilizes his available inventory (products in stock) to satisfy customers' demand; as a result, Agent 1 suffers either an inventory holding cost (for excess inventory) or a backorder cost (for excess demand). Then, Agent 1 decides his desired inventory level for round $t+1$ and orders from Agent 2. Next, Agent 2 handles the order from Agent 1, suffers inventory holding costs or backorder costs, decides her base-stock level for round $t+1$, and places an order from an external source (assumed to have infinite inventory). The two agents' orders will arrive at the beginning of the next round. The optimal policy with known demand distributions is known as the *base-stock* policy for both agents (Clark and Scarf, 1960; Federgruen and Zipkin, 1984; Chen and Zheng, 1994). Specifically, a base-stock policy keeps a fixed base-stock level s over all time periods, meaning that if the inventory level (on-hand inventory minus the backlogged ordered) at the beginning of a period is below s, an order will be placed to bring the inventory level to s; otherwise, no order is placed.

There are recently works extending the classic inventory control problem with known demand distribution to the one with *unknown* distribution (Levi et al., 2007; Huh and Rusmevichientong, 2009; Huh et al., 2011; Levi et al., 2015; Zhang et al., 2018; Chen et al., 2020, 2021; Chen and Chao,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

2020; Ding et al., 2021). However, these works consider the single-agent case, instead of the two-echelon case. In this work, we aim at extending the classic two-echelon stochastic inventory planning problem to an online setup with an *unknown* demand distribution $\mathcal{D}$. In addition, we consider the *nonperishable* setting in which any leftover inventory will be carried over to the next round; as a result, the inventory level at the beginning of the next round can not be lower than the inventory level at the end of current round. The performance is measured by i) regret, the difference between their total loss and that of the best base-stock policy in hindsight; ii) last-iterate convergence to the best base-stock policy for both agents.

It is important to note that Agent 2 only observes orders from Agent 1 and does not necessarily receive the same demand information as Agent 1 does. In addition, in our problem formulation, Agent 1's inventory will be impacted by Agent 2's shortages. Specifically, when Agent 2 does not have enough inventory to fill Agent 1's order, we assume that Agent 2 cannot expedite to meet the shortfall, and this shortfall will cause a partial shipment to Agent 1, which implies that Agent 1 may not achieve his desired inventory level at the beginning of each round. This model with a known demand distribution is first examined in (Cachon and Zipkin, 1999).

We consider two different decision-making settings: *centralized* and *decentralized* settings. The centralized setting takes the perspective of a central planner who decides both agents' desired inventory level at each round in order to minimize the total loss of the entire supply chain. A more interesting and realistic setting concerns a decentralized structure in which the two agents independently decide their own desired inventory level at each round to minimize their own costs, which often results in poor performance of the supply chain (i.e., the optimal base-stock level for each agent may not be the one that achieves minimal overall loss). To achieve the optimal supply chain performance under the decentralized setting, as discussed in previous works (i.e. (Cachon, 2003)), some mechanism concerns contractual arrangement or corporate rules, such as rules for sharing the holding costs and backorder costs, accounting methods, and/or operational constraints. A *contract* transfers the loss between the two agents such that each agent's objective is aligned with the supply chain's objective. However, as far as we know, this is only discussed under known demand distribution. Thus, we extend the results to the online setting with an unknown demand distribution and design learning algorithms to achieve the optimal supply chain performance.

1.1 Techniques and Results

Techniques. Our problem has three salient features that are different from the classic stochastic online convex optimization problem. First, as will be shown in Section 3, the overall loss function is not convex with respect to both agents' inventory decisions, meaning that we can not directly apply online convex optimization algorithms to this problem. Second, due to the multi-echelon nature of the supply chain, Agent 2's input information is dependent on the information generated by Agent 1, which can be non-stochastic. Third, in the nonperishable setting, each agent's inventory level at the beginning of the next round *can not* be lower than the inventory level at the end of the current round, which implies that the desired inventory level may not be always achievable.

To address the first challenge, we introduce an augmented loss function upon which is convex and we are able to perform online convex optimization algorithms. To address the second and the third challenge, our algorithm for both agents has the low-switching property, which only updates the strategy $\mathcal{O}(\log T)$ times. This makes Agent 2's input information almost the same as the realized demand at each round. For Agent 1, as he can always observe the true realized demand at each round in both centralized and decentralized setting, he makes his inventory decision based on the empirical demand distribution, which is updated $\mathcal{O}(\log T)$ times during the process. For Agent 2, in the centralized and the decentralized setting, our algorithm is a variant of Online Gradient Descent (OGD) and Online Newton Step (ONS) (Hazan et al., 2007), respectively. Both of the algorithms have the important low-switching property, which only updates the strategy $\mathcal{O}(\log T)$ times while at the same time achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ and $\mathcal{O}(\log^2 T)$ regret respectively. We remark that our variant of ONS algorithm achieves $\mathcal{O}(\log^2 T)$ regret, even when the loss function is not strongly convex but satisfies a certain property.

Our results. In the centralized setting, we design an algorithm which achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ expected regret and last-iterate convergence to the optimal base-stock policy with rate $\widetilde{\mathcal{O}}(1/\sqrt{T})$ for Agent 1 and $\widetilde{\mathcal{O}}(T^{-1/4})$ for Agent 2. In the decentralized setting, we design a novel contract mechanism and also learning algorithms for both agents, which lead to convergence to both agents' global optimal base-stock policy with the same rate as the centralized setting. In addition, our algorithm guarantees that Agent 1 has $\widetilde{\mathcal{O}}(T^{3/4})$ individual expected regret and Agent 2 has $\mathcal{O}(\log^3 T)$ individual expected regret. Moreover, the expected regret with respect to the overall loss is bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$, which is the same as the one in the centralized setting. Table 1 shows a summary of our results.[1] We also implement our algorithms, and the empirical results validate the effectiveness of our algorithms (see Appendix D). To the best of our knowledge, our work is the first one considering the two-echelon stochastic inventory planning problem in the online setup with unknown demand distribution.

[1] All our expected regret bound can be extended to high-probability regret bound with an $\mathcal{O}(\sqrt{T\log(1/\delta)})$ overhead by applying standard Azuma's inequality.

Table 1: Summary of our results. "Centralized" and "Decentralized" represent the centralized and decentralized settings, respectively. The definitions of Reg_T, $\text{Reg}_{T,1}$ and $\text{Reg}_{T,2}$ are introduced in Section 3. "Convergence for Agent 1" and "Convergence for Agent 2" represent the convergence rate to Agent 1 and Agent 2's optimal inventory level, respectively.

Setting	Reg_T	$\text{Reg}_{T,1}$	$\text{Reg}_{T,2}$	Convergence for Agent 1	Convergence for Agent 2
Centralized	$\widetilde{\mathcal{O}}(\sqrt{T})$	N/A	N/A	$\widetilde{\mathcal{O}}(1/\sqrt{T})$	$\widetilde{\mathcal{O}}(T^{-1/4})$
Decentralized	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(T^{3/4})$	$\widetilde{\mathcal{O}}(\log^3 T)$	$\widetilde{\mathcal{O}}(1/\sqrt{T})$	$\widetilde{\mathcal{O}}(T^{-1/4})$

1.2 Related Works

There is a vast body of SCM literature on achieving the optimal supply chain performance in the decentralized setting (Lariviere, 1999; Tsay et al., 1999; Cachon, 2003; Chen, 2003) concerning coordination with contract design and information sharing. In this body of literature, there is a line of works based on multi-echelon decentralized inventory models, which are closely related to our study, including (Lee and Whang, 1999; Cachon and Zipkin, 1999; Lee et al., 2000; Porteus, 2000; Watson and Zheng, 2005; Shang et al., 2009). However, these works all assume that the demand distribution is known (at least to the downstream agent).

More recently, there has been growing interest in single-agent inventory control problems with unknown demand distribution (Levi et al., 2007; Huh and Rusmevichientong, 2009; Huh et al., 2011; Levi et al., 2015; Zhang et al., 2018; Chen et al., 2020, 2021; Chen and Chao, 2020; Ding et al., 2021). In particular, (Huh and Rusmevichientong, 2009) achieves $\mathcal{O}(\log T)$ regret in the perishable setting and $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the nonperishable setting using online gradient descent method. (Ding et al., 2021) further extends the results to the feature-based setting. The nonparametric approach of this line of works is fundamentally different from the conventional inventory control models in which the inventory manager knows the demand distribution (see, e.g., (Zipkin, 2000; Snyder and Shen, 2019) for comprehensive reviews of the conventional inventory models); however, unlike the conventional inventory theory which has been extended from the single-echelon problems to multi-echelon problems, little has been done for the multi-echelon problems under unknown demand distributions, and we aim to fill in this gap.

The other relevant line of works is online convex optimization. (Zinkevich, 2003) shows that OGD algorithm achieves $\mathcal{O}(\sqrt{T})$ expected regret bound for general convex functions. If the loss functions are exp-concave, (Hazan et al., 2007) shows that ONS achieves $\mathcal{O}(\log T)$ expected regret bound. Both algorithms change their decision at every round. On the other hand, Sherman and Koren (2021) proposes a lazy version of OGD, which changes its decision only $\mathcal{O}(\log T)$ times and still achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ (or $\mathcal{O}(\log^2 T)$) regret when the loss functions are stochastically generated and convex (or strongly convex). In our problem, it turns out to be crucial to apply an algorithm with a small number of switches, and our algorithm generalizes the idea of (Sherman and Koren, 2021) to the ONS algorithm to achieve $\mathcal{O}(\log T)$ switches and $\widetilde{\mathcal{O}}(1)$ regret for a larger class of functions including strong convex functions.

2 Preliminary

Notations. For a positive integer n, denote $[n]$ to be the set $\{1, 2, \ldots, n\}$. For conciseness, we hide polynomial dependence on the problem-dependent constants in the $\mathcal{O}(\cdot)$ notation and only show the dependence on the horizon T. $\widetilde{\mathcal{O}}(\cdot)$ further hides the poly-logarithmic dependency on T. Define $(x)^+ \triangleq \max\{x, 0\}$ and $(x)^- \triangleq \max\{-x, 0\}$. $\|x\|$ denotes the Euclidean norm of x.

Throughout this work, we make the following two mild assumptions on the demand distribution $\mathcal{D}$. These mild assumptions are also made in (Chen et al., 2020).

Assumption 1. The demand distribution $\mathcal{D}$ is supported on $[d, D]$ where $D > d > 0$.

Assumption 2. The image of the density function of $\mathcal{D}$, $\phi(\cdot)$, lies in $[\gamma, \Gamma]$ where $\Gamma > \gamma > 0$.

Under the above demand assumptions, we consider the following model in the two-echelon inventory planning problem, which is first considered in (Cachon and Zipkin, 1999). Our goal is to find the best base-stock policy.

We first introduce the cost function under a fixed base-stock policy. In this model, we assume that Agent 2's inventory shortage will cause delayed (by one round) shipment and shortfalls at Agent 1 while Agent 2's orders will always be satisfied as we assume that the external source has infinite inventory. In addition, for unfilled demand for Agent 1, there is a backorder cost shared by the two agents, αp_1 for Agent 1 and $(1-\alpha)p_1$ for Agent 2, where α is the negotiated cost sharing parameters via contractual arrangements. The inventory holding cost per unit for Agent 1 and Agent 2 is h_1 and h_2 respectively.

Now we are ready to define the loss function for Agent 1 and Agent 2 respectively. Specifically, Agent 1's loss function is formulated as follows. Define

$$G_1(y) = h_1 \mathbb{E}_{x\sim\mathcal{D}}[(y-x)^+] + \alpha p_1 \mathbb{E}_{x\sim\mathcal{D}}[(y-x)^-],$$

which is Agent 1's expected sum of the holding and backorder costs per round with *unlimited supply* under base-stock level y. Since the actual supply to Agent 1 is limited by Agent 2's available inventory, according to (Cachon and Zipkin, 1999), Agent 1's expected sum of the holding and backorder costs per period is defined as

$$H_1(s_1, s_2) \triangleq \Phi(s_2)G_1(s_1) + \int_{s_2}^{D} G_1(s_1 + s_2 - x)\phi(x)dx,$$

where $\Phi(\cdot)$ is the cumulative density function of $\mathcal{D}$. The first term is Agent 1's costs when Agent 2 has sufficient inventory to satisfy Agent 1's order (i.e., Agent 1's inventory level can be brought up to s_1), while the second term is the cost when Agent 2 does not have enough inventory to satisfy Agent 1's order, meaning that Agent 2's shortfall is $x - s_2$ and Agent 1's inventory can only be brought up to $s_1 + s_2 - x$.

For Agent 2, define

$$G_2(y) = (1 - \alpha)p_1\mathbb{E}_{x\sim\mathcal{D}}[(y - x)^-],$$

which is the expected backorder cost per period incurred by Agent 2 due to Agent 1's shortages. Then, the expected backorder cost incurred by Agent 2 is

$$\Phi(s_2)G_2(s_1) + \int_{s_2}^{D} G_2(s_1 + s_2 - x)\phi(x)dx.$$

The first term is the backorder cost incurred by Agent 2 due to Agent 1's shortfalls when the Agent 1's inventory level is s_1, while the second term is the backorder cost incurred by Agent 2 when Agent 1's inventory level is $s_1 - (x - s_2) < s_1$. As can be seen, Agent 2's shortages $(x - s_2)$ will cause insufficient supply to Agent 1, which, in turn, will be detrimental to Agent 2 herself when Agent 1 is out of stock due to the insufficient supply. Therefore, Agent 2's loss function is the sum of the expected backorder cost and the expected holding cost, which is defined as

$$H_2(s_1, s_2) \triangleq h_2\mathbb{E}_{x\sim\mathcal{D}}[(s_2 - x)^+] + \Phi(s_2)G_2(s_1) + \int_{s_2}^{D} G_2(s_1 + s_2 - x)\phi(x)dx.$$

We also define the sum of both agents loss as $H(s_1, s_2) \triangleq H_1(s_1, s_2) + H_2(s_1, s_2)$ and $G(s) \triangleq G_1(s) + G_2(s)$.

Online Inventory Control In this work, we study this conventional model in an online learning setting that proceeds in T rounds. Before the game starts, both Agent 1 and Agent 2 order an initial inventory level $s_{1,1}$ and $s_{1,2}$. Then, for each round $t \in [T]$:

- at the start of round t, both agents' orders arrive. The current inventory level for Agent 1 and Agent 2 reaches to $\widehat{s}_{t,1}$ and $\widehat{s}_{t,2}$;
- external demand d_t occurs at Agent 1's level where d_t is drawn from the unknown demand distribution $\mathcal{D}$. In this step, Agent 1 suffers from some inventory holding cost or backorder cost. Define the inventory level for Agent 1 after demand as $\widetilde{s}_{t,1}$. This value can be negative as we assume backlogged orders;
- Agent 1 decides his desired inventory level at the next round $s_{t+1,1}$, which leads to a demand for Agent 2: $o_t = (s_{t+1,1} - \widetilde{s}_{t,1})^+$;
- Agent 2 receives the demand o_t from Agent 1, and the inventory level after demand is $\widetilde{s}_{t,2}$. Note that, in general, Agent 2 only knows o_t instead of the real demand d_t. Agent 2 then suffers some inventory holding cost or backorder cost;
- Agent 2 decides her desired inventory level for the next round $s_{t+1,2}$ and orders $o'_t = (s_{t+1,2} - \widetilde{s}_{t,2})^+$ from some external source.

We remark that the dynamic of the inventory for Agent 1 and Agent 2 are different. As we assume that the external source has infinite inventory, Agent 2's order can always be satisfied and we have the following dynamic for $\widetilde{s}_{t,2}$ and $\widehat{s}_{t+1,2}$:

$$\widetilde{s}_{t,2} = \widehat{s}_{t,2} - o_t, \quad \widehat{s}_{t+1,2} = \widetilde{s}_{t,2} + o'_t. \tag{1}$$

However, as Agent 2 may have delayed shipment when she does not have enough inventory, Agent 1's dynamic is defined as follows. Define the delayed shipment of Agent 2 as $(o_{t-1} - \widehat{s}_{t-1,2})^+$, which will arrive after Agent 1 has served the demand d_t. This means that

$$\widetilde{s}_{t,1} = \widehat{s}_{t,1} - d_t + (o_{t-1} - \widehat{s}_{t-1,2})^+,$$
$$\widehat{s}_{t+1,1} = \widetilde{s}_{t,1} + \min\{\widehat{s}_{t,2}, o_t\}.$$

The specific costs suffered by the two agents in each step, as well as their objectives will be discussed in detail in Section 3.2.

3 Main Results

3.1 Centralized Setting

In this section, we start from considering the centralized setting of our model where there is a central planner who decides both agents' strategy simultaneously. Define the loss suffered by the learner at round t as follows:

$$\widetilde{H}_t = h_1(\widehat{s}_{t,1} - d_t)^+ + p_1(\widehat{s}_{t,1} - d_t)^- + h_2(\widehat{s}_{t,2} - o_t)^+,$$

and the benchmark as the expected loss suffered by the best base-stock policy: $H(s_1^*, s_2^*)$ where $(s_1^*, s_2^*) = \operatorname{argmin}_{s_1,s_2} H(s_1, s_2)$. The regret is defined as the difference between the sum of the learners' total loss and the loss

Algorithm 1 Central Planner for Coupling Model

Input: An instance of stochastic OGD $\mathcal{A}$ (Algorithm 2).
Initialize: Arbitrary empirical cumulative density function $\widehat{\Phi}_0(\cdot)$. Epoch length $L_1 = 1$. $\tau = 1$.
for $m = 1, 2, \ldots$ **do**
  Define epoch $I_m = \{\tau, \tau+1, \ldots, \tau+L_m-1\}$.
  Set $s_{m,1} = \widehat{\Phi}_{m-1}^{-1}(\frac{h_2+p_1}{h_1+p_1})$.
  Receive $s_{m,2}$ from $\mathcal{A}$.
  while $\tau \in I_m$ **do**
    Decide the desired inventory level for both agents: $s_{m,1}$ for Agent 1 and $s_{m,2}$ for Agent 2.
    Receive the realized demand d_τ, $\tau \leftarrow \tau + 1$.
  end
  Collect $\mathcal{D}_m = \{d_{t'}\}_{t' \in I_m}$; define $\widehat{\Phi}_m(x) = \frac{1}{L_m}\sum_{\tau \in I_m} \mathbb{I}\{d_\tau \le x\}$ and also the inverse function $\widehat{\Phi}_m^{-1}(z) = \min\{x : \widehat{\Phi}(x) \ge z\}$; send $\widehat{\Phi}_m(x)$ and $\mathcal{D}_m$ to $\mathcal{A}$; and set $L_{m+1} = 2L_m$.
end

of the best base-stock policy, which is formally written as follows:

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\widetilde{H}_t - H(s_1^*, s_2^*))\right].$$

Compared with the standard online convex optimization problem (Zinkevich, 2003), in which at each round the loss suffered in each round is a convex function of the current decision, our problem has two main difficulties. First, in our problem, the loss of the algorithm in each round depends not only on the current decided order-up-to level $s_{t,1}$ and $s_{t,2}$, but also the past decisions $s_{\tau,1}$ and $s_{\tau,2}$ for $\tau \in [t]$ as we consider the non-perishable setting, meaning that the ordered inventories can not be discarded. Second, even under fixed base-stock policy, the loss function $H(s_1, s_2)$ is *not jointly convex* (also not convex in s_2). In the following, we show how we handle these two difficulties respectively.

To deal with the first difficulty, our first key observation is that if both agents' decision $s_{t,1}$ and $s_{t,2}$ are changing very infrequently, then the loss of the algorithm at each round is almost equivalent to $\widehat{H}_t(s_1, s_2)$, which is defined as:

$$\begin{aligned}&\widehat{H}_t(s_1, s_2) \\ &\triangleq h_1(\mathring{s}_{t,1} - d_t)^+ + p_1(\mathring{s}_{t,1} - d_t)^- + h_2(s_2 - d_t)^+,\end{aligned} \tag{2}$$

where $\mathring{s}_{t,1} = s_1$ if $s_2 > d_{t-1}$ and $\mathring{s}_{t,1} = s_1 + s_2 - d_{t-1}$ if $s_2 \le d_{t-1}$. Note that this loss function is a stochastic function and is only dependent on the current decision variables (s_1, s_2).

To see why the loss function at round t can be almost written as $\widehat{H}_t(s_{t,1}, s_{t,2})$ if both agents' decisions do not change very frequently, we first point out the two differences between $\widetilde{H}_t$ and $\widehat{H}_t(s_1, s_2)$. First, as agents can not discard the inventories that have been ordered, the true inventory level $\widehat{s}_{t,1}$ at the beginning of round t may not be the desired inventory level $s_{t,1}$ when Agent 2 does not have an inventory shortage, or $s_{t,1}+s_{t,2}-d_{t-1}$ when Agent 2 has an inventory shortage. Similarly, Agent 2's true inventory level $\widehat{s}_{t,2}$ may not be her desired inventory level $s_{t,2}$. Recall that $\widehat{s}_{t,1}$ and $\widehat{s}_{t,2}$ are used in defining $\widetilde{H}_t$. Second, in $\widehat{H}_t(s_{t,1}, s_{t,2})$, the demand of Agent 2 equals to d_t, while in the definition of $\widetilde{H}_t$, the demand for Agent 2 is the order amount o_t from Agent 1.

Fortunately, these two differences can both be properly handled by a low-switching algorithm. Specifically, suppose that both agents' desired inventory levels are kept the same: $s_{t,1} = s_1'$ and $s_{t,2} = s_2'$ for all t in some time period $[t_0, t_0 + L]$. Then, we can show that

$$\widehat{s}_{t,1} = \begin{cases} s_1', & \text{if } s_2' > d_{t-1}, \\ s_1' + s_2' - d_{t-1}, & \text{otherwise}, \end{cases} \tag{3}$$

$$\widehat{s}_{t,2} = s_2', \tag{4}$$

$$o_t = d_t, \tag{5}$$

except for at most $\Theta(1)$ rounds at the beginning of the period $[t_0, t_0 + L]$, making $\widetilde{H}_t = \widehat{H}_t(s_{t,1}, s_{t,2})$ for all the rest of the rounds. This is because Equation (3) does not hold only when $\widetilde{s}_{t-1,1} > s_{t,1} = s_1'$, which can only happen for $\Theta(1)$ rounds as the demand at each round is strictly larger than 0 according to Assumption 1. Then, as $o_t = (s_{t+1,1} - \widetilde{s}_{t,1})^+ = (s_{t+1,1} - \widehat{s}_{t,1} + d_t)^+$, when $\widehat{s}_{t,1} = s_1' = s_{t+1,1}$, we know that $o_t = d_t$. Similarly, as $\widehat{s}_{t,2} \ne s_{t,2}$ only happens when $\widetilde{s}_{t-1,2} > s_{t,2} = s_2'$, and $o_t = d_t$ is strictly positive after $\Theta(1)$ rounds, we know that $\widehat{s}_{t,2} = s_{t,2} = s_2'$ after another $\Theta(1)$ rounds. This argument is formally summarized below and proven in Appendix A.

Lemma 3.1. *In round t_0, suppose that Agent 1 and Agent 2's desired inventory level for the following L rounds is s_1' and s_2'. Then, for some $t_1 = \Theta(1)$, it holds that for all $t \in [t_0 + t_1, t_0 + L]$, $\widehat{s}_{t,2} = s_2'$, $o_t = d_t$. In addition, $\widehat{s}_{t,1} = s_1'$ if $s_2' > d_{t-1}$ and $\widehat{s}_{t,1} = s_1' + s_2' - d_{t-1}$ otherwise. Consequently, it holds that $\widetilde{H}_t = \widehat{H}_t(s_1', s_2')$ during $t \in [t_0 + t_1, t_0 + L]$.*

In addition, as proven in Lemma A.2 in the appendix, it holds that $\mathbb{E}[\widehat{H}_t(s_1, s_2)] = H(s_1, s_2)$. This reduces our problem to optimizing over the stochastic loss $\widehat{H}_t(s_1, s_2)$ with infrequent changes.

Next, we show how we handle the second difficulty, which is the issue of non-convexity of our loss function. Our second key observation is that with direct calculation, one can show that the optimal base-stock policy of Agent 1 has the close form: $s_1^* = \Phi^{-1}(\frac{h_2+p_1}{h_1+p_1})$ and that $H(s_1^*, s_2)$ is now convex in s_2; see Lemma A.3 for a formal proof. Ideally, if we set Agent 1's desired inventory level to be s_1^*, then we are able to apply gradient descent method to learn the

best base-stock policy for Agent 2. However, we do not have the knowledge of the true demand distribution. Therefore, our solution is to construct an empirical cumulative density function $\widehat{\Phi}_L(\cdot)$ for the demand distribution during the learning process where $\widehat{\Phi}_L(\cdot)$ is constructed by using L i.i.d. demand samples. Then, let Agent 1's desired inventory level be $s_{L,1} = \widehat{\Phi}_L^{-1}(\frac{h_2+p_1}{h_1+p_1})$.

However, the expected loss function $H(s_{L,1}, s_2)$ may still not be convex in s_2 due to the approximation error of the empirical cumulative density function. To handle this issue, we introduce the following *augmented* loss function:

$$H'_L(s_{L,1}, s_2) \triangleq H(s_{L,1}, s_2) + (h_1+p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\int_0^{s_2}\Phi(x)dx, \tag{6}$$

where $C_1 > 0$ is some universal constant specified in Lemma A.4. We show in Lemma A.5 that $H'_L(s_{L,1}, s_2)$ is indeed convex in s_2 with high probability.

Combining the above augmented loss function design with the idea of having a low-switching algorithm, we design our centralized algorithm Algorithm 1 as follows. The algorithm goes in epochs with exponentially increasing lengths, meaning that the number of epochs is only $\mathcal{O}(\log T)$. At the beginning of the m-th epoch I_m, both agents decide a fixed desired inventory level for this epoch. Specifically, Agent 1 chooses his level as $s_{m,1} = \widehat{\Phi}_{m-1}^{-1}(\frac{h_2+p_1}{h_1+p_1})$ (Line 2) where $\widehat{\Phi}_{m-1}(\cdot)$ is the empirical cumulative density function constructed by the observed demand samples within the previous epoch I_{m-1}. Standard concentration (Lemma A.4) shows that $s_{m,1}$ converges to s_1^*. As discussed before, with this choice of $s_{m,1}$, the loss function $H(s_{m,1}, s_2)$ may still not convex in s_2. According to Equation (6), with a slight abuse of notation, we introduce the augmented loss function for Agent 2 at epoch m:

$$H'_m(s_{m,1}, s_2) \triangleq H(s_{m,1}, s_2) + (h_1+p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L_{m-1}}}\int_0^{s_2}\Phi(x)dx, \tag{7}$$

where L_m is the length of epoch I_m and C_1 is the same as the one in Equation (6). As proven in Lemma A.5, with high probability, $H'_m(s_{m,1}, s_2)$ is convex in s_2, which enables us to apply stochastic OGD to minimize this (unknown) loss function via demands received in the previous epoch and output the average iterate as the desired inventory level for Agent 2. The full pseudo code is shown in Algorithm 2.

This concludes our algorithm design for the centralized setting of our model. Note that as the number of epoch is $\mathcal{O}(\log T)$ and both agents pick a fixed desired inventory level within each epoch, there are at most $\Theta(\log T)$ number of rounds such that Equation (3), Equation (4) and Equation (5) do not hold. In addition, as the epoch length L_{m-1} gets longer, $H'_m(s_{m,1}, s_2)$ will get closer to the true loss function $H(s_{m,1}, s_2)$. Combined with the fact that $s_{m,1}$ is converging to s_1^* when m grows, the output of Algorithm 2, which is the average iterate of stochastic OGD, will converge to s_2^* as well. Moreover, it can be shown that Algorithm 1 achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. See the formal statement below, the proof in Appendix A, and empirical results in Section 4 and Appendix D.

Algorithm 2 Centralized Algorithm for Agent 2

Input: A set of demand value $\mathcal{D} = \{d_1, \ldots, d_L\}$, empirical cumulative density function $\widehat{\Phi}_L(x) = \frac{1}{L}\sum_{i=1}^{L}\mathbb{I}\{d_i \le x\}$, learning rate $\eta > 0$ and failure probability δ.
Initialize: Set $s_{1,2} \le D - \frac{h_2}{\Gamma(h_2+p_1)} = s_{\max}$ arbitrarily.
Set $s_1 = \widehat{\Phi}_L^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)$.
for $t = 1, 2, \ldots, L$ **do**
  $s_{t+1,2} = \min\{s_{\max}, \max\{0, (s_{t,2} - \eta \cdot m_t)\}\}$, where $m_t = (h_1+p_1)\mathbb{I}\{\widehat{s}_{t,1} \ge d_t\} - p_1 + h_2\mathbb{I}\{s_{t,2} \ge d_t\} + C_1(h_1+p_1)\sqrt{\frac{\log(TD/\delta)}{L}} \cdot \widehat{\Phi}_L(s_{t,2})$ and $\widehat{s}_{t,1} = s_1$ if $d_{t-1} \le s_{t,2}$ and $\widehat{s}_{t,1} = s_1 + s_{t,2} - d_{t-1}$ otherwise.
end
return $\bar{s}_{L,2} = \frac{1}{L}\sum_{\tau=1}^{L} s_{\tau,2}$.

Theorem 3.2. *Algorithm 1 guarantees that with probability at least $1-2\delta$, the strategy converges to the optimal base-stock policy with the following rate:*

$$|s_{M,1} - s_1^*| \le \mathcal{O}\left(\sqrt{\log(T/\delta)/T}\right),$$
$$|s_{M,2} - s_2^*| \le \mathcal{O}\left(T^{-1/4}\log^{1/4}(T/\delta)\right),$$

with $M = \mathcal{O}(\log T)$ the number of epochs. Picking $\delta = 1/T^2$, Algorithm 1 guarantees that $\mathbb{E}[\mathrm{Reg}_T] \le \widetilde{\mathcal{O}}(\sqrt{T})$.

3.2 Decentralized Setting with Contracts

In this section, we consider how both agents learn the optimal base-stock policy (s_1^*, s_2^*) in the decentralized setting where each agent decides their desired inventory level independently. As shown by (Cachon and Zipkin, 1999), in the offline setting with *known* demand distribution, to guarantee that each agent's own optimal inventory level matches the overall optimal level, a *contract* is needed to reallocate the inventory holding and backorder costs between the two agents through linear payments. This contract mechanism is widely used in SCM (Cachon and Zipkin, 1999; Lee and Whang, 1999). Specifically, we design a contract between the two agents, which sets $\alpha = 1$, meaning that Agent 1 is responsible for all penalty costs due to his shortages, and decides a coefficient ω, which is the cost that Agent 2 needs to compensate Agent 1 for each unsatisfied order requested by Agent 1.

Therefore, we define the loss suffered by Agent 1 and Agent 2 at round t as follows:

$$\widetilde{H}^{\mathrm{c}}_{t,1} \triangleq h_1(\widehat{s}_{t,1} - d_t)^+ + p_1(\widehat{s}_{t,1} - d_t)^- - \omega_t(\widehat{s}_{t,2} - o_t)^-,$$
$$\widetilde{H}^{\mathrm{c}}_{t,2} \triangleq h_2(\widehat{s}_{t,2} - o_t)^+ + \omega_t(\widehat{s}_{t,2} - o_t)^-,$$

where ω_t is the contract coefficient agreed by both agents at round t. The benchmark is the loss suffered by the best base-stock policy for each agent defined as follows:

$$\mathring{H}^{\mathrm{c}}_{t,1}(\mathring{s}^*_1)$$
$$\triangleq h_1(\mathring{s}^*_{t,1} - d_t)^+ + p_1(\mathring{s}^*_{t,1} - d_t)^- - \omega_t(\widehat{s}_{t,2} - d_t)^-,$$
$$\mathring{H}^{\mathrm{c}}_{t,2}(\mathring{s}^*_2) \triangleq h_2(\mathring{s}^*_2 - o_t)^+ + \omega_t(\mathring{s}^*_2 - o_t)^-,$$

where $\mathring{s}^*_1 = \operatorname{argmin}_{s_1} \mathbb{E}[\sum_{t=1}^T \mathring{H}^{\mathrm{c}}_{t,1}(s_1)]$, $\mathring{s}^*_2 = \operatorname{argmin}_{s_2} \mathbb{E}[\sum_{t=1}^T \mathring{H}^{\mathrm{c}}_{t,2}(s_2)]$ and $\mathring{s}^*_{t,1}$ is Agent 1's inventory level at the beginning of round t if he uses the base-stock policy $\mathring{s}^*_1$. Note that when Agent 1 keeps a fixed base-stock policy, it holds that $o_t = d_t$ for all $t \in [T]$. The expected regret for each agent is defined as

$$\mathbb{E}\left[\mathrm{Reg}_{T,1}\right] \triangleq \mathbb{E}\left[\sum_{t=1}^T \widetilde{H}^{\mathrm{c}}_{t,1} - \sum_{t=1}^T \mathring{H}^{\mathrm{c}}_{t,1}(\mathring{s}^*_1)\right],$$
$$\mathbb{E}\left[\mathrm{Reg}_{T,2}\right] \triangleq \mathbb{E}\left[\sum_{t=1}^T \widetilde{H}^{\mathrm{c}}_{t,2} - \sum_{t=1}^T \mathring{H}^{\mathrm{c}}_{t,2}(\mathring{s}^*_2)\right].$$

Now we introduce the design of our algorithm in the decentralized setting. Similar to the centralized setting, in order to make sure that the loss function for each of the agent is almost only dependent on the current desired inventory level, the algorithm we design still satisfies that both agents do not update their desired inventory level very frequently, making Equation (3), Equation (4) and Equation (5) hold almost all the time. First, we introduce Agent 1's algorithm. As Agent 1 can still observe the true demand at each round, he is able to apply the same process as shown in Algorithm 1. Specifically, Agent 1 still breaks the total horizon into $\mathcal{O}(\log T)$ epochs with exponentially increasing length and chooses his desired inventory level to be $\widehat{\Phi}^{-1}_{m-1}(\frac{h_2+p_1}{h_1+p_1})$ at each epoch I_m to converge to s^*_1.

Next, we consider the design of Agent 2's algorithm. Although Agent 2 can also run a variant of Algorithm 2 as Agent 1 only changes his desired level $\mathcal{O}(\log T)$ times, making $o_t = d_t$ except for $\mathcal{O}(\log T)$ rounds, Agent 2 will suffer a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret due to the approximation error of the cumulative density function. To achieve a better regret bound for Agent 2, note that if Agent 2 applies a low-switching algorithm and in addition, $\omega_t = \omega, o_t = d_t$ for all $t \in [T]$, then Agent 2's loss can almost be written as $\widehat{H}^{\mathrm{c},\omega}_{t,2}(s)$ defined as follows:

$$\widehat{H}^{\mathrm{c},\omega}_{t,2}(s) \triangleq h_2\left(s - d_t\right)^+ + \omega\left(s - d_t\right)^-, \tag{8}$$

as we know that there are only few rounds such that $\widehat{s}_{t,2} \neq s$ according to Lemma 3.1. In addition, direct calculation shows that $\operatorname{argmin}_s \mathbb{E}_{d_t \sim \mathcal{D}}[\widehat{H}^{\mathrm{c},\omega}_{t,2}(s)] = \Phi^{-1}(\frac{\omega}{\omega+h_2})$.

Now, we focus on regret minimization with respect to $\widehat{H}^{\mathrm{c}}_{t,2}$. Our key observation here is that this loss function is not only convex, but also satisfy the so-called *Bernstein Stochastic Gradients* property that allows faster learning. Specifically, we prove in Lemma B.1 that $\widehat{H}^{\mathrm{c},\omega}_{t,2}$ satisfies the following property (with a specific choice of $B > 0$).[2]

Property 1. Let $\mathcal{F}$ be a distribution over a class of convex functions $f : \mathcal{X} \mapsto \mathbb{R}^d$. We say $\mathcal{F}$ satisfies B-Bernstein condition with $B > 0$ if for all $x \in \mathcal{X}$, we have

$$(x - x^*)^\top \mathbb{E}_{f\sim\mathcal{F}}\left[\nabla f(x) \nabla f(x)^\top\right](x - x^*)$$
$$\leq B(x - x^*)^\top \mathbb{E}_{f\sim\mathcal{F}}\left[\nabla f(x)\right],$$

where $x^* = \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}_{f\sim\mathcal{F}}[f(x)]$.

As shown by Van Erven and Koolen (2016), there exist learning algorithms that achieve $\mathcal{O}(\log T)$ regret bound when facing a sequence of loss functions f_t, each drawn independently from $\mathcal{F}$ that satisfies Property 1. As a simplification, we show that the classic ONS algorithm (Algorithm 4 shown in Appendix B.1) with a proper learning rate is already able to achieve $\mathcal{O}(\log T)$ regret in this case; see Theorem B.3.

However, classic ONS changes its decision at each round. Taking inspiration from (Sherman and Koren, 2021), we indeed succeed in designing a low-switching variant of ONS that changes its decision only $\widetilde{\mathcal{O}}(1)$ times while still ensuring $\widetilde{\mathcal{O}}(1)$ regret under Property 1. Specifically, our algorithm (Algorithm 5 shown in Appendix B.1) divides the total horizon into $\mathcal{O}(\log T)$ epochs with exponentially increasing lengths. While in each round it still performs the ONS update, the actual decision is only updated at the beginning of each epoch, which is set to the average of all previous ONS decisions. It is clear that our algorithm only switches its decision $\mathcal{O}(\log T)$ times. More importantly, we show that the price for the regret is only an extra $\mathcal{O}(\log T)$ factor, leading to an overall $\mathcal{O}(\log^2 T)$ regret; see Theorem B.4 for the formal statement.[3] In additional to enjoying $\mathcal{O}(\log^2 T)$ regret, our algorithm in fact also ensures the last-iterate convergence to the global optimal solution of the expected loss function. This is due to the strong convexity of $\mathbb{E}[\widehat{H}^{\mathrm{c},\omega}_{t,2}(s)]$ and the fact that the last-iterate is the average of ONS updates in the previous epochs. We summarize the results in Lemma B.5 and the full proof is presented in Appendix B.4. This means that if Agent 1 changes his desired

[2] We show in Lemma B.2 that $\widehat{H}^{\mathrm{c},\omega}_{t,1}(s_1)$ satisfies Property 1 even when the demand is discrete.

[3] In fact, one can show that a lazy version of OGD also achieves similar guarantees. However, this heavily relies on a continuous demand distribution, while ONS works even for discrete demands (see Footnote 2). Note that a discrete demand distribution is common in applications since demands usually come in a batch.

Algorithm 3 Protocol among Agent 1, Agent 2 and contract maker

Input: Failure probability δ. Initial epoch length $L_1 = 16\rho \triangleq 256(h_2+p_1)^4 h_2^{-4} C_3^4 \log^4(T/\delta)$, where $C_3 > 0$ is a constant defined in Equation (22).
Initialize: $s_{0,2}$ and $\widehat{\Phi}_0(\cdot)$ arbitrarily. $\tau = 1$.
for $m = 1, 2, \ldots$ **do**
1. Define epoch $I_m = \{\tau, \tau+1, \ldots, \tau+L_m-1\}$.
2. Agent 1 and Agent 2 receive the contract coefficient ω_m from contract maker using Algorithm 2.
3. Agent 1 chooses $s_{m,1} = \widehat{\Phi}_{m-1}^{-1}(\frac{h_2+p_1}{h_1+p_1})$. Initialize an instance $\mathcal{A}$ of Algorithm 5 with $\eta = \max\{\omega_m^2, h_2^2\}/(\gamma(h_2+\omega_m))$, $\varepsilon = 1/T$ and initial decision $s_{\tau-1,2}$.
   while $\tau \in I_m$ **do**
4. Agent 1 receives demand d_τ.
5. Agent 2 receives her ordered products from the outer resource and Agent 1 receives his unsatisfied demand at $\tau-1$ and his inventory level goes to $\widetilde{s}_{\tau,1}$.
6. Agent 1 decides his desired inventory level at $\tau+1$ to be $s_{m,1}$, sends the order $o_\tau = \max\{s_{m,1} - \widetilde{s}_{\tau,1}, 0\}$ to Agent 2, and suffers loss $\widetilde{H}^{\mathrm{c}}_{\tau,1}$.
7. Agent 2 sends the loss function $f_\tau(x) = h_2\mathbb{I}\{x \geq o_\tau\} + \omega_m\mathbb{I}\{x \leq o_\tau\}$ to $\mathcal{A}$, and suffers loss $\widetilde{H}^{\mathrm{c}}_{\tau,2}$.
8. Agent 2 sets her own desired inventory level to be $s_{\tau,2}$, which is the output of $\mathcal{A}$.
   end
9. Agent 1 collects $\mathcal{D}_m = \{d_{t'}\}_{t' \in I_m}$, computes $\widehat{\Phi}_m(x) = \frac{1}{L_m}\sum_{\tau \in I_m}\mathbb{I}\{d_\tau \leq x\}$ and also the inverse function $\widehat{\Phi}_m^{-1}(z) = \min\{x : \widehat{\Phi}_m(x) \geq z\}$. Set $L_{m+1} = 2L_m$.

end

inventory level $\mathcal{O}(\log T)$ times and Agent 2 applies Algorithm 5 on her loss with $\omega_t = \omega$ for all $t \in [T]$, she will suffer $\mathcal{O}(\log^2 T)$ regret and converge to base-stock policy $\Phi^{-1}(\frac{\omega}{\omega+h_2})$. Therefore, if $\omega_t = \omega^* = {h_2\Phi(s_2^*)}/{1-\Phi(s_2^*)}$ for all $t \in [T]$, then applying Algorithm 5, Agent 2 will converge to s_2^*.

Thus, it remains to figure out how to learn ω^* for the contract maker. We design an algorithm which updates the contract coefficient ω at the beginning of each epoch I_m of Agent 1. With a slight abuse of notation, let ω_m be the contract during epoch I_m. As the contract maker can observe the realized demand as well as both agents cost parameters, at the beginning of epoch I_m, we run Algorithm 2 to obtain an (imaginary) inventory level $s'_{m,2}$ for Agent 2 given L_{m-1} demand samples collected during epoch I_{m-1}, and then calculate ω_m following $\frac{h_2\Phi(s_2^*)}{1-\Phi(s_2^*)}$ with s_2^* replaced by $s'_{m,2}$ and Φ replaced by the empirical cumulative density function. The algorithm is shown in Algorithm 6 and deferred to Appendix B.5. The following lemma shows that given enough samples from the demand distribution, with high probability, Algorithm 6 outputs a contract coefficient ω that is very close to the ideal coefficient ω^*. The proof can be found in Appendix B.6.

Lemma 3.3. *Let $L \geq \rho$ where ρ is defined in Algorithm 3. Given L i.i.d samples $\{d_i\}_{i=1}^L$ from the demand distribution $\mathcal{D}$, with probability at least $1-\delta$, Algorithm 6 guarantees that i) $|\omega - \omega^*| \leq \mathcal{O}(L^{-1/4}\log(T/\delta))$, where $\omega^* = \frac{h_2\Phi(s_2^*)}{1-\Phi(s_2^*)}$; and ii) $\omega \in [0, h_2+p_1+\mathcal{O}(L^{-1/4}\log(T/\delta))]$.*

Given this lemma, we now provide an overview of our algorithm (Algorithm 3). It proceeds in epochs with exponentially increasing lengths again. At the beginning of epoch I_m, both agents receive a contract coefficient ω_m calculated via Algorithm 6 (Line 2). Then Agent 1 decides his desired inventory level to be $\widehat{\Phi}_{m-1}^{-1}(\frac{h_2+p_1}{h_1+p_1})$ based on his observed demand in epoch I_{m-1}. Agent 2 then initializes an instance of Algorithm 5 with the decision from the last round of the previous epoch as the initial decision (Line 3), and uses this instance to decide her inventory level for this epoch (Line 7 and Line 8). The following theorem shows that Algorithm 3 guarantees the convergence to the offline optimal solution as well as sublinear regret for both agents. The proof is deferred to Appendix B.7.

Theorem 3.4. *Algorithm 3 guarantees that with probability at least $1-3\delta$,*

$$|s_{M,1} - s_1^*| \leq \mathcal{O}\left(\sqrt{\log(T/\delta)/T}\right),$$
$$|s_{M,2} - s_2^*| \leq \mathcal{O}\left(T^{-1/4}\log(T/\delta)\right),$$

where $M = \mathcal{O}(\log T)$ the total number of epochs. Picking $\delta = 1/T^2$, Algorithm 3 guarantees that $\mathbb{E}[\mathrm{Reg}_{T,1}] \leq \widetilde{\mathcal{O}}(T^{3/4})$ and $\mathbb{E}[\mathrm{Reg}_{T,2}] \leq \mathcal{O}(\log^3 T)$.

Finally, we show that Algorithm 3 also guarantees that the regret with respect to the sum of both agents' losses is bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$, which is the same as the one obtained in the centralized setting. Note that by directly using the regret guarantees, the convergence of both agents decisions in Theorem 3.4, and the lipschitzness of the loss function, the overall regret of the loss sum can only be bounded by $\widetilde{\mathcal{O}}(T^{3/4})$. In order to improve the overall regret from $\widetilde{\mathcal{O}}(T^{3/4})$ to $\widetilde{\mathcal{O}}(\sqrt{T})$, we need to apply a refined analysis on Agent 2's decision sequence; see the following see Appendix B.8 for the proof of the following theorem. Empirical results shown in Section 4 and Appendix D also support our theoretical statements.

Theorem 3.5. *Picking $\delta = 1/T^3$, Algorithm 3 guarantees that $\mathbb{E}[\mathrm{Reg}_T] \leq \widetilde{\mathcal{O}}(\sqrt{T})$.*

4 Experiment

In this section, we show the empirical performance of our designed algorithms. We implement the Explore-then-Exploit algorithm, in which both agents first spend $\lceil T^{\frac{2}{3}} \rceil$ rounds

(h_1, h_2, p_1)	Distribution	Agent 1 Loss			Agent 2 Loss			Loss sum				
		Explore-then-Exploit	Algorithm 3	Imp (%)	Explore-then-Exploit	Algorithm 3	Imp (%)	Explore-then-Exploit	Algorithm 1	Imp (%)	Algorithm 3	Imp (%)
$(0.3, 0.1, 0.5)$	Gaussian	1.6376(0.0072)	1.4689(0.0061)	10.30%	0.371(0.0025)	0.3027(0.0062)	18.42%	2.0086(0.0067)	1.7656(0.0066)	12.10%	1.7716(0.0044)	11.80%
	Exponential	2.1009(0.013)	1.8795(0.011)	10.54%	0.466(0.0058)	0.3838(0.0112)	17.63%	2.5669(0.0101)	2.2452(0.0077)	12.53%	2.2633(0.0053)	11.83%
	Uniform	2.0384(0.0091)	1.9304(0.007)	5.30%	0.4518(0.0031)	0.3929(0.0058)	13.05%	2.4902(0.0084)	2.313(0.0136)	7.12%	2.3233(0.0048)	6.70%
$(0.4, 0.25, 0.6)$	Gaussian	2.2086(0.0061)	1.9779(0.0141)	10.45%	0.6244(0.0031)	0.4151(0.0134)	33.52%	2.833(0.0061)	2.3872(0.0059)	15.74%	2.393(0.003)	15.53%
	Exponential	2.9418(0.0091)	2.652(0.0161)	9.85%	0.58(0.0082)	0.2998(0.0152)	48.32%	3.5218(0.0053)	2.9647(0.0171)	15.82%	2.9518(0.0054)	16.19%
	Uniform	2.7329(0.0094)	2.5752(0.0082)	5.77%	0.7418(0.0055)	0.5544(0.0089)	25.27%	3.4747(0.009)	3.1228(0.0115)	10.13%	3.1295(0.005)	9.93%
$(0.5, 0.35, 1.5)$	Gaussian	3.4114(0.0103)	3.2144(0.0113)	5.77%	0.9955(0.0033)	0.7046(0.0107)	29.22%	4.4069(0.0095)	3.9115(0.0078)	11.24%	3.919(0.0058)	11.07%
	Exponential	4.6274(0.0112)	4.3911(0.0147)	5.11%	0.9569(0.008)	0.5583(0.013)	41.66%	5.5844(0.0102)	4.9476(0.0181)	11.40%	4.9494(0.0081)	11.37%
	Uniform	4.1014(0.0123)	4.0023(0.0122)	2.41%	1.3272(0.0055)	1.0825(0.0099)	18.44%	5.4286(0.0108)	5.0661(0.0198)	6.68%	5.0849(0.0068)	6.33%
$(0.6, 0.4, 2.0)$	Gaussian	4.1805(0.0117)	3.9578(0.01)	5.33%	1.217(0.0049)	0.89(0.0082)	26.86%	5.3975(0.0106)	4.8422(0.01)	10.29%	4.8478(0.0082)	10.18%
	Exponential	5.6462(0.0148)	5.3838(0.0163)	4.65%	1.2314(0.0074)	0.7841(0.0137)	36.32%	6.8776(0.0134)	6.1566(0.0109)	10.48%	6.1679(0.0099)	10.32%
	Uniform	5.0052(0.016)	4.9028(0.0104)	2.05%	1.6306(0.0048)	1.3579(0.008)	16.72%	6.6358(0.0166)	6.2478(0.0213)	5.85%	6.2607(0.0093)	5.65%

Table 2: Empirical results of our algorithm for coupling model with $T = 400000$. The contract coefficient at each round is generated by Algorithm 6. Each algorithm is processed over 32 trials of demand sequences drawn from the three distributions. The mean and the standard deviation of the time-averaged loss over the 32 trials of format "mean (std)" are shown in the table. "Imp (%)" shows the amount of improvement of our algorithm compared to the baseline Explore-then-Exploit algorithm. The results show that our proposed algorithm (Algorithm 3) in the decentralized setting outperforms the vanilla Explore-then-Exploit algorithm in the perspective of each agent's individual loss. From the perspective of overall loss, Algorithm 1 performs the best and Algorithm 3 still performs better than the vanilla Explore-then-Exploit algorithm, showing the effectiveness of our proposed algorithms.

picking a fixed base-stock policy which is uniformly randomly drawn from $[d, D]$. Then, both agents use the collected realized demand samples to construct the empirical density function of the demand and switch to the optimal base-stock policy with respect to the empirical demand distribution for the remaining $T - \lceil T^{\frac{2}{3}} \rceil$ rounds. The demand distributions we constructed are as follows:

- Gaussian distribution $\mathcal{N}(10, 5)$ clipped on support $[0.5, 20]$;
- Uniform distribution over $[0.5, 20]$;
- Exponential distribution with mean 10 clipped on support $[0.5, 20]$.

We set the number of round $T = 400000$ and choose the cost configuration to be $(h_1, h_2, p_1) = (0.3, 0.1, 0.5), (0.4, 0.25, 0.6), (0.5, 0.35, 1.5), (0.6, 0.4, 2.0)$. For each demand distribution, 32 trials are processed and we calculate the mean and the standard deviation of each agent's individual loss with the contract coefficient generated by Algorithm 6 over the 32 trials. The time-averaged losses for each agents with mean and standard deviation are shown in Table 2. From Table 2, we can observe that from the perspective of each agent's individual loss, our proposed decentralized algorithm (Algorithm 3) outperforms the vanilla Explore-then-Exploit algorithm. Specifically, our algorithm suffers about $10\% \sim 40\%$ less per-round loss compared to the one suffered by the vanilla Explore-then-Exploit algorithm. In the sense of overall loss, our proposed centralized algorithm (Algorithm 1) performs the best among the three algorithms and the performance of the decentralized algorithm (Algorithm 3) is still $10\% \sim 20\%$ better than the one of vanilla Explore-then-Exploit, which shows the effectiveness of our proposed algorithms. More experiment results are shown in Appendix D.

5 Conclusion and Future Directions

In contrast to the classic offline two-echelon stochastic inventory planning problem with known distribution studied in the SCM literature, we consider the problem with an unknown demand distribution in an online setting, which is more realistic and, as far as we know, not studied before. We consider the model formulation introduced in (Cachon and Zipkin, 1999) under both the centralized and decentralized setting, and prove both regret guarantees and convergence to the offline optimal base-stock policy. While we assume that the true demand is observable even when it exceeds the current inventory level, a more challenging setting is the censored demand setting where only the amount of the satisfied demand is available. Extending our results to the censored demand and unobserved lost sales setting appears to require new ideas.

Acknowledgements

MZ and HL are supported by NSF Award IIS-1943607.

References

Gérard P. Cachon and Paul H. Zipkin. Competitive and cooperative inventory policies in a two-stage supply chain. *Management Science*, 45(7):936–953, 1999. ISSN 00251909, 15265501.

David Simchi-Levi, Philip Kaminsky, and Edith Simchi-Levi. *Designing and managing the supply chain: con-*

cepts, strategies and case studies. McGraw-Hill/Irwin, 1999.

Andrew J Clark and Herbert Scarf. Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4):475–490, 1960.

Awi Federgruen and Paul Zipkin. Computational issues in an infinite-horizon, multi-echelon inventory model. *Operations Research*, 32(4):818–836, 1984.

Fangruo Chen and Yu-Sheng Zheng. Lower bounds for multi-echelon stochastic inventory systems. *Management Science*, 40(11):1426–1443, 1994.

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

Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009.

Woonghee Tim Huh, Retsef Levi, Paat Rusmevichientong, and James B Orlin. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4):929–941, 2011.

Retsef Levi, Georgia Perakis, and Joline Uichanco. The data-driven newsvendor problem: new bounds and insights. *Operations Research*, 63(6):1294–1306, 2015.

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

Weidong Chen, Cong Shi, and Izak Duenyas. Optimal learning algorithms for stochastic inventory systems with random capacities. *Production and Operations Management*, 29(7):1624–1649, 2020.

Boxiao Chen, Xiuli Chao, and Cong Shi. Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research*, 46(2):726–756, 2021.

Boxiao Chen and Xiuli Chao. Dynamic inventory control with stockout substitution and demand learning. *Management Science*, 66(11):5108–5127, 2020.

Jingying Ding, Woonghee Tim Huh, and Ying Rong. Feature-based nonparametric inventory control with censored demand. *Available at SSRN 3803777*, 2021.

Gérard P Cachon. Supply chain coordination with contracts. *Handbooks in operations research and management science*, 11:227–339, 2003.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Martin A Lariviere. Supply chain contracting and coordination with stochastic demand. In *Quantitative models for supply chain management*, pages 233–268. Springer, 1999.

Andy A Tsay, Steven Nahmias, and Narendra Agrawal. Modeling supply chain contracts: A review. *Quantitative models for supply chain management*, pages 299–336, 1999.

Fangruo Chen. Information sharing and supply chain coordination. *Handbooks in operations research and management science*, 11:341–421, 2003.

Hau Lee and Seungjin Whang. Decentralized multi-echelon supply chains: Incentives and information. *Management science*, 45(5):633–640, 1999.

Hau L Lee, Kut C So, and Christopher S Tang. The value of information sharing in a two-level supply chain. *Management science*, 46(5):626–643, 2000.

Evan L Porteus. Responsibility tokens in supply chain management. *Manufacturing & Service Operations Management*, 2(2):203–219, 2000.

Noel Watson and Yu-Sheng Zheng. Decentralized serial supply chains subject to order delays and information distortion: Exploiting real-time sales data. *Manufacturing & Service Operations Management*, 7(2):152–168, 2005.

Kevin H Shang, Jing-Sheng Song, and Paul H Zipkin. Coordination mechanisms in decentralized serial inventory systems with batch ordering. *Management Science*, 55 (4):685–695, 2009.

Paul Herbert Zipkin. *Foundations of inventory management*. McGraw-Hill, 2000.

Lawrence V Snyder and Zuo-Jun Max Shen. *Fundamentals of supply chain theory*. John Wiley & Sons, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning*, pages 928–936, 2003.

Uri Sherman and Tomer Koren. Lazy oco: Online convex optimization on a switching budget. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3972–3988. PMLR, 15–19 Aug 2021.

Tim Van Erven and Wouter M Koolen. Metagrad: Multiple learning rates in online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

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

A Omitted proofs in Section 3.1

In this section, we show the omitted proofs in the centralized setting in Section 3.

First, we prove Lemma 3.1, which shows that if both agents keep picking the same desired inventory level $(s_{t,1}, s_{t,2}) = (s_1', s_2')$ for a period of rounds, then there are at most $\Theta(1)$ rounds such that Equation (3), Equation (4) and Equation (5) do not hold. For completeness, we restate Lemma 3.1 as follows:

Lemma A.1 (Restatement of Lemma 3.1)**.** *In round t_0, suppose that Agent 1 and Agent 2's desired inventory level for the following L rounds is s_1' and s_2'. Then, for some $t_1 = \Theta(1)$, it holds that for all $t \in [t_0 + t_1, t_0 + L]$, $\widehat{s}_{t,2} = s_2'$, $d_t = o_t$. In addition, $\widehat{s}_{t,1} = s_1'$ if $s_2' > d_{t-1}$ and $\widehat{s}_{t,1} = s_1' + s_2' - d_{t-1}$ otherwise. Consequently, it hold that $\widetilde{H}_t = \widehat{H}_t(s_1', s_2')$ for all $t \in [t_0 + t_1, t_0 + L]$.*

Proof. Let $\tau^* = \operatorname{argmin}_{\tau \in [L]} \{o_{t_0+\tau} > 0\}$. Next, we first show that $\{o_{t_0+\tau}\}_{\tau=\tau^*+1}^{L-1} = \{d_{t+\tau}\}_{\tau=\tau^*+1}^{L-1}$ and $\widetilde{s}_{t,1} = s_1' - d_t$ for all $t \in \{\tau^* + 1, \tau^* + 2, \ldots, L-1\}$. If $s_1' \geq \widetilde{s}_{t_0,1}$, then we have $o_{t_0} = s_1' - \widetilde{s}_{t_0,1} \geq 0$. As Agent 1 will receive the unsatisfied orders from Agent 2 before Agent 1 makes the order in the next round, at round $t+1$, Agent 1's inventory before ordering is $\widetilde{s}_{t_0+1} = s_1' - d_{t_0+1}$, which means that $o_{t_0+1} = s_1' - \widetilde{s}_{t_0+1} = d_{t_0+1}$. Repeating the above process shows that for all $t \in [t_0 + 1, t_0 + L]$, we have $o_t = d_t$ and $\widetilde{s}_{t,1} = s_1' - d_t$.

On the other hand, if $s_1' < \widetilde{s}_{t_0,1}$, then we have $o_{t_0} = 0$ as Agent 1 can not discard the inventory. According to Assumption 1, we have $d_{t_0+1} \geq d$ and $\widetilde{s}_{t_0+1,1} \leq \widetilde{s}_{t_0,1} - d$. Therefore, within at most constant $t_2 = \mathcal{O}(1)$ number of rounds, we have $\widetilde{s}_{t+t_2,1} \geq s_1'$. Then following the analysis in the first case proves that during $t \in [t_0 + t_2, t_0 + L]$, we have $o_t = d_t$ and $\widetilde{s}_{t,1} = s_1' - d_t$.

Next, we show that $\widehat{s}_{t,2} = s_2$ after constant number of rounds. Specifically, if $\widehat{s}_{t_0,2} \leq s_2'$, then at round $t_0 + 1$, we have $\widehat{s}_{t_0+1,2} = s_2'$. Otherwise, note that when $t' \geq t_0 + t_2$, $o_{t'} = d_{t'} \geq d$. Therefore, after at most constant $t_3 = \mathcal{O}(1)$ number of rounds, we have $\widehat{s}_{t_0+t_2+t_3,2} \leq s_2'$, meaning that $\widehat{s}_{t,2} = s_2'$ for all $t \in [t_0 + t_2 + t_3 + 1, t_0 + L]$.

Finally, we show that $\widehat{s}_{t,1} = s_1'$ if $s_2' > d_{t-1}$ and $\widehat{s}_{t,1} = s_1' + s_2' - d_{t-1}$ otherwise after constant number of rounds. As shown above, when $t \geq t_0 + t_2 + t_3 + 1$, we know that $\widehat{s}_{t,2} = s_2'$, $\widetilde{s}_{t,1} = s_1' - d_t$ and $o_t = d_t$. According to the dynamic of $\widehat{s}_{t,1}$, we know that

$$\widehat{s}_{t+1,1} = \widetilde{s}_{t,1} + \min\{\widehat{s}_{t,2}, o_t\} = s_1' - d_t + \min\{s_2', d_t\} = \begin{cases} s_1' + s_2' - d_t & \text{if } s_2' < d_t, \\ s_1' & \text{otherwise.} \end{cases}$$

Therefore, setting $t_1 = t_2 + t_3 + 1 = \mathcal{O}(1)$ finishes the proof of the first statement. The second statement holds for $t \in [t_0 + t_1, t_0 + L]$ according to the definition of $\widetilde{H}_t$ and $\widehat{H}_t(s_1', s_2')$. □

Next, we show that stochastic loss function defined in Equation (2) is an unbiased loss estimator of the expected loss $H(s_1, s_2)$.

Lemma A.2. $\mathbb{E}\left[\widehat{H}_t(s_1, s_2)\right] = H(s_1, s_2)$*, for all $t \in [T]$, where $\widehat{H}_t(s_1, s_2)$ is defined in Equation (2).*

Proof. According to the definition of $\widetilde{H}_t$, we know that

$$\begin{aligned}
\mathbb{E}\left[\widehat{H}_t(s_1, s_2)\right] &= \mathbb{E}\left[h_1(\mathring{s}_{t,1} - d_t)^+ + p_1(\mathring{s}_{t,1} - d_t)^- + h_2(s_2 - d_t)^+\right] \\
&= \mathbb{E}\left[\left(h_1(s_1 - d_t)^+ + p_1(s_1 - d_t)^-\right) \cdot \mathbb{I}\left\{s_1 \in [s_1 - d_{t-1}, s_1 - d_{t-1} + s_2]\right\}\right] + \mathbb{E}\left[h_2(s_2 - d_t)^+\right]. \\
&\quad + \mathbb{E}\left[\left(h_1(s_1 + s_2 - d_{t-1} - d_t)^+ + p_1(s_1 + s_2 - d_{t-1} - d_t)^-\right) \cdot \mathbb{I}\left\{s_1 \geq s_1 - d_{t-1} + s_2\right\}\right] \\
&= \mathbb{E}\left[\left(h_1(s_1 - d_t)^+ + p_1(s_1 - d_t)^-\right) \cdot \mathbb{I}\left\{s_2 \geq d_{t-1}\right\}\right] + \mathbb{E}\left[h_2(s_2 - d_t)^+\right]. \\
&\quad + \mathbb{E}\left[\left(h_1(s_1 + s_2 - d_{t-1} - d_t)^+ + p_1(s_1 + s_2 - d_{t-1} - d_t)^-\right) \cdot \mathbb{I}\left\{s_2 \leq d_{t-1}\right\}\right] \\
&= \Phi(s_2) \cdot G(s_1) + \int_{s_2}^{D} G(s_1 + s_2 - u)\phi(u)du = H(s_1, s_2).
\end{aligned}$$

□

The next lemma shows that the optimal solution (s_1^*, s_2^*) of $H(s_1, s_2)$ satisfies that $s_1^* = \Phi^{-1}(\frac{h_2+p_1}{h_1+p_1})$ and $H(s_1^*, s_2)$ is convex in s_2.

Lemma A.3. *Let $(s_1^*, s_2^*) = \operatorname{argmin}_{s_1,s_2} H(s_1, s_2)$. Then it holds that $s_1^* = \Phi^{-1}(\frac{h_2+p_1}{h_1+p_1})$ and $H(s_1^*, s_2)$ is convex in s_2, where $\Phi(\cdot)$ is the cumulative density function of demand distribution $\mathcal{D}$.*

Proof. By definition of $H(s_1, s_2)$, it holds that

$$\begin{aligned} H(s_1, s_2) &= h_2\mathbb{E}_{x\sim\mathcal{D}}[(s_2 - x)^+] + \Phi(s_2)G(s_1) + \int_{s_2}^{D} G(s_1 + s_2 - u)\phi(u)du \\ &= h_2\mathbb{E}_{x\sim\mathcal{D}}[(s_2 - x)^+] + \Phi(s_2)G(s_1) + \int_{0}^{D-s_2} G(s_1 - u)\phi(u + s_2)du. \end{aligned}$$

Taking gradient over s_1 and s_2 respectively, we know that

$$\begin{aligned} \nabla_{s_1} H(s_1, s_2) &= (p_1 + h_1)\Phi(s_2)\left(\Phi(s_1) - \frac{p_1}{p_1 + h_1}\right) + (p_1 + h_1)\int_0^{D-s_2}\left(\Phi(s_1 - u) - \frac{p_1}{p_1 + h_1}\right)\phi(u + s_2)du \\ &= (p_1 + h_1)\Phi(s_2)\Phi(s_1) - p_1 + (p_1 + h_1)\int_0^{D-s_2}\Phi(s_1 - u)\phi(u + s_2)du, \\ \nabla_{s_2} H(s_1, s_2) &= h_2\Phi(s_2) + \phi(s_2)G(s_1) + \int_0^{D-s_2} G(s_1 - u)d\phi(u + s_2) - G(s_1 + s_2 - D)\phi(D) \\ &= h_2\Phi(s_2) + \phi(s_2)G(s_1) + [G(s_1 - u)\phi(u + s_2)]_0^{D-s_2} \\ &\quad + \int_0^{D-s_2}\phi(u + s_2)(h_1 + p_1)\left(\Phi(s_1 - u) - \frac{p_1}{p_1 + h_1}\right) - G(s_1 + s_2 - D)\phi(D) \\ &= h_2\Phi(s_2) - p_1(1 - \Phi(s_2)) + (h_1 + p_1)\int_0^{D-s_2}\Phi(s_1 - u)\phi(u + s_2)du. \end{aligned} \tag{9}$$

Setting the two gradients to be 0, we obtain that

$$\begin{aligned} &\Phi(s_2)\Phi(s_1) + \int_0^{D-s_2}\Phi(s_1 - u)\phi(u + s_2)du = \frac{p_1}{p_1 + h_1}, \\ &0 = \nabla_{s_2} H(s_1, s_2) = (h_2 + p_1)\Phi(s_2) - p_1 - (h_1 + p_1)\Phi(s_2)\Phi(s_1) + p_1 = \Phi(s_2)\left[(p_1 + h_2) - (h_1 + p_1)\Phi(s_1)\right]. \end{aligned}$$

Therefore, we have $s_1^* = \Phi^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)$ and s_2^* satisfies that

$$(h_2 + p_1)\Phi(s_2^*) + (p_1 + h_1)\int_0^{D-s_2^*}\Phi(s_1^* - u)\phi(u + s_2^*)du = p_1.$$

Replacing s_1 by s_1^* in Equation (9) and taking gradient over s_2, we obtain that

$$\begin{aligned} \nabla_{s_2}^2 H(s_1^*, s_2) &= (h_2 + p_1)\phi(s_2) + (h_1 + p_1)\cdot\left(\int_0^{D-s_2}\phi(s_2 + u)\phi(s_1^* - u)du - \Phi(s_1^*)\phi(s_2)\right) \\ &\geq \left((h_2 + p_1) - (h_1 + p_1)\Phi(s_1^*)\right)\phi(s_2) = 0, \end{aligned} \tag{10}$$

showing that $H(s_1^*, s_2)$ is convex in s_2.

□

The next lemma follows by the standard concentration inequality, showing that the gap between $\Phi(s_{m,1})$ and $\Phi(s_1^*)$ is bounded by $\widetilde{\mathcal{O}}(L_{m-1}^{-1/2})$ with high probability.

Lemma A.4. *Let $d_1, \ldots, d_T$ be T i.i.d. samples from distribution $\mathcal{D}$ which satisfies Assumption 1. Let the empirical density distribution constructed by $\{d_i\}_{i=1}^L$ as $\widehat{\Phi}_L(\cdot)$ defined as Equation (36) and the inverse of the empirical density function $\widehat{\Phi}_L^{-1}(\cdot)$ as defined in Equation (37). Let $s_{L,1} = \widehat{\Phi}_L^{-1}(\frac{h_2+p_1}{h_1+p_1})$. Then, with probability at least $1 - \delta$, for all $L \in [T]$,*

$$|\Phi(s_{L,1}) - \Phi(s_1^*)| \leq C_1\sqrt{\frac{\log(TD/\delta)}{L}},$$

where $C_1 > 0$ is some universal constant.

Proof. Direct calculation shows that for all $L \in [T]$,

$$|\Phi(s_{L,1}) - \Phi(s_1^*)| \le \Gamma\,|s_{L,1} - s_1^*| \le \Gamma \left|\widehat{\Phi}_L^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right) - \Phi^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)\right| \le C_0\sqrt{\frac{\log(TD/\delta')}{L}},$$

where the first inequality is by Assumption 2, the second inequality is by definition of $s_{L,1}$ and s_1^*, and the last inequality holds with probability $1-\delta$ by Lemma C.3. □

The next lemma shows that with probability at least $1-\delta$, our constructed augmented loss function in Equation (6):

$$H'_L(s_{L,1}, s_2) = H(s_{L,1}, s_2) + (h_1+p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\int_0^{s_2}\Phi(x)dx$$

is convex in s_2 for all $L \in [T]$, where $s_{L,1} = \widehat{\Phi}_L^{-1}(\frac{h_2+p_1}{h_1+p_1})$ and $C_1 > 0$ is defined in Lemma A.4. This ensures that using (stochastic) online gradient descent with respect to $H'_m(s_{m,1}, \cdot)$ defined in Equation (7) achieves sublinear regret.

Lemma A.5. *With probability at least $1-\delta$, $H'_L(s_{L,1}, s_2) \triangleq H(s_{L,1}, s_2) + (h_1+p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\int_0^{s_2}\Phi(x)dx$ is convex in s_2, for all $L \in [T]$. Consequently, with probability at least $1-\delta$, for all $m \in [M]$, $H'_m(s_{m,1}, s_2)$ defined in Equation (7) is convex, where $M = \mathcal{O}(\log T)$ is the number of epochs.*

Proof. According to the definition of $H'_L(s_{L,1}, s_2)$, we obtain that the second-order gradient on the second parameter s_2 equals to

$$\begin{aligned}
g_{22} &= \nabla_{s_2}^2 H'_L(s_{L,1}, s_2) \\
&= [(h_2+p_1) - (h_1+p_1)\Phi(s_{L,1})]\cdot\phi(s_2) + (h_1+p_1)\int_{s_2}^{D}\phi(s_{L,1}+s_2-u)\phi(u)du \\
&\quad + C_1(h_1+p_1)\sqrt{\frac{\log(TD/\delta)}{L}}\phi(s_2) \\
&\ge \left[(h_2+p_1) - (h_1+p_1)\left(\Phi(s_{L,1}) - C_1\sqrt{\frac{\log(TD/\delta)}{L}}\right)\right]\phi(s_2) \\
&\ge [(h_2+p_1) - (h_1+p_1)\Phi(s_1^*)]\,\phi(s_2) = 0,
\end{aligned}$$

where the last inequality holds for all $L \in [T]$ with probability at least $1-\delta$ according to Lemma A.4. Therefore, we know that with probability at least $1-\delta$, $H'_L(s_{L,1}, s_2)$ is a convex function for all $L \in [T]$, meaning that $H'_m(s_{m,1}, s_2)$ is also convex in s_2 for all $m \in [M]$ with probability at least $1-\delta$, where $M = \mathcal{O}(\log T)$ is the total number of epochs. □

In the next lemma, we show that Algorithm 2, which applies stochastic online gradient descent with respect to s_2 on the augmented loss function, enjoys average-iterate convergence to the optimal solution.

Lemma A.6. *Given L i.i.d samples $\{d_i\}_{i=1}^L$ from the demand distribution $\mathcal{D}$ with $L \ge \log(TD/\delta)$. Let $s_{L,2}$ be the inventory level output by Algorithm 2 $\mathcal{A}$. Then with probability at least $1-\delta$, we have $|s_{L,2} - s_2^*| \le \mathcal{O}(L^{-1/4}\log^{1/4}(TD/\delta))$.*

Proof. According to Lemma A.4, we know that with probability at least $1-\delta$, for any $L \in [T]$,

$$|\Phi(s_{L,1}) - \Phi(s_1^*)| \le \Gamma\,|s_{L,1} - s_1^*| \le \Gamma \left|\widehat{\Phi}_L^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right) - \Phi^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)\right| \le C_1\sqrt{\frac{\log(TD/\delta)}{L}}, \tag{11}$$

Direct calculation shows that the gradient of $H'_L(s_{L,1}, s_2)$ with respect to s_2 is as follows:

$$\nabla_{s_2} H'_L(s_{L,1}, s_2) = \nabla_{s_2} H(s_{L,1}, s_2) + (h_1+p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\Phi(s_2). \tag{12}$$

As shown in Lemma A.2, we know that

$$H(s_{L,1}, s_2) = \mathbb{E}\left[h_1(\widehat{s}_{L,1} - d_t)^+ + p_1(\widehat{s}_{L,1} - d_t)^- + h_2(s_2 - d_t)^+\right],$$

where $\widehat{s}_{L,1} = s_{L,1}$ if $s_2 \le d_{t-1}$ and $\widehat{s}_{L,1} = s_{L,1} + s_2 - d_{t-1}$ otherwise. Therefore, an unbiased estimator of the first term of the right hand side of Equation (12) is as follows:

$$(h_1 + p_1)\mathbb{I}\{\widehat{s}_{L,1} \ge d_t\} - p_1 + h_2\mathbb{I}\{s_2 \ge d_t\}.$$

For the second term, as $\widehat{\Phi}_L(\cdot)$ is an unbiased estimator of $\Phi(\cdot)$, we know that

$$(h_1 + p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\Phi(s_2) = (h_1 + p_1)C_1\sqrt{\frac{\log(TD/\delta)}{L}}\mathbb{E}\left[\widehat{\Phi}_L(s_2)\right].$$

Therefore, we can indeed construct an unbiased estimator of the true gradient $\nabla_{s_2} H'_L(\widehat{s}_{L,1}, s_2)$ and run stochastic online gradient descent. Specifically, as shown in Algorithm 2, let

$$m_t = (h_1 + p_1)\mathbb{I}\{\widehat{s}_{L,1} \ge d_t\} - p_1 + h_2\mathbb{I}\{s_{t,2} \ge d_t\} + C_1(h_1 + p_1)\sqrt{\frac{\log(TD/\delta)}{L}}\widehat{\Phi}_L(s_{t,2}).$$

Then based on the above calculation, we know that $\mathbb{E}[m_t] = \nabla_{s_2} H'_L(\widehat{s}^*_1, s_{t,2})$. Moreover, as $L \ge \log(TD/\delta)$, we know that $|m_t| \le \max\{h_1, p_1\} + C_1(h_1 + p_1) = \mathcal{O}(1)$. According to classic online gradient descent analysis (e.g. Theorem 3.1.1 in (Hazan et al., 2016)), Algorithm 2 guarantees that for any $s_2 \le D - \frac{h_2}{\Gamma(h_2+p_1)}$,

$$\mathbb{E}\left[\sum_{\tau=1}^{L} H'_L(s_{L,1}, s_{\tau,2}) - \sum_{\tau=1}^{L} H'_L(s_{L,1}, s_2)\right] \le \mathcal{O}(\sqrt{L}), \tag{13}$$

where we omit all the problem-dependent constants here.

Next, we show that $s^*_2 \le D - \frac{h_2}{\Gamma(h_2+p_1)}$. From the optimality condition of s^*_1 and s^*_2 and Equation (9), we know that

$$\begin{aligned}
s^*_1 &= \Phi^{-1}\left(\frac{h_2 + p_1}{h_1 + p_1}\right), \\
0 = \nabla_{s_2} H(s^*_1, s^*_2) &= (h_2 + p_1)\Phi(s^*_2) - p_1 + (p_1 + h_1)\int_0^{D - s^*_2} \Phi(s^*_1 - u)\phi(u + s^*_2)du \\
&\ge (h_2 + p_1)\Phi(s^*_2) - p_1.
\end{aligned}$$

Therefore, it holds that

$$s^*_2 \le \Phi^{-1}\left(\frac{p_1}{h_2 + p_1}\right). \tag{14}$$

Furthermore, as $\phi(x) \in [\gamma, \Gamma]$ for all $x \in [d, D]$,

$$\Gamma(D - s^*_2) \ge \Phi(D) - \Phi(s^*_2) \ge 1 - \frac{p_1}{h_2 + p_1} = \frac{h_2}{h_2 + p_1} \implies s^*_2 \le D - \frac{h_2}{\Gamma(h_2 + p_1)}.$$

Therefore, according to the $\max\{h_1, p_1\}$-Lipschitzness of $H'_L(s_1, s_2)$ in s_1, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=1}^{L} H'_L(s^*_1, s_{\tau,2}) - \sum_{\tau=1}^{L} H'_L(s^*_1, s_2)\right] \\
&\le \mathbb{E}\left[\sum_{\tau=1}^{L} H'_L(s_{L,1}, s_{\tau,2}) - \sum_{\tau=1}^{L} H'_L(s_{L,1}, s_2)\right] + \mathcal{O}\left(\sqrt{L\log(TD/\delta)}\right) \\
&\qquad \text{(Lipschitzness of } H'_L(s_1, s_2) \text{ and Equation (11))} \\
&\le \mathcal{O}(\sqrt{L\log(TD/\delta)}). \qquad \text{(by Equation (13))}
\end{aligned}$$

Choosing $s_2 = s_2^*$, $\delta = \frac{1}{T^2}$ and using the definition of $H'_L(s_1, s_2)$, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=1}^{L} H(s_1^*, s_{\tau,2}) - \sum_{t=1}^{L} H(s_1^*, s_2^*)\right] \\
&\leq \mathbb{E}\left[\sum_{\tau=1}^{L} H'_L(s_1^*, s_{\tau,2}) - \sum_{\tau=1}^{L} H'_L(s_1^*, s_2^*)\right] + 2C_1 L(h_1+p_1)\mu \cdot \sqrt{\frac{\log(TD/\delta)}{L}} + \mathcal{O}(1) \\
&\leq \mathbb{E}\left[\sum_{\tau=1}^{L} H'_L(s_1^*, s_{\tau,2}) - \sum_{\tau=1}^{L} H'_L(s_1^*, s_2^*)\right] + 2C_1(h_1+p_1)\mu\sqrt{L\log(TD/\delta)} + \mathcal{O}(1) \\
&\leq \mathcal{O}(\sqrt{L\log T}),
\end{aligned} \tag{15}$$

where $\mu = \mathbb{E}_{x\sim\mathcal{D}}[x] \leq D$ and $\mathcal{O}(\cdot)$ hides all problem-dependent constants.

Moreover, note that $H(s_1^*, s_2)$ is σ_2''-strongly convex in $s_2 \leq D - \frac{h_2}{\Gamma(h_2+p_1)}$ as according to Equation (10),

$$\begin{aligned}
\nabla^2 H(s_1^*, s_2) &= [(h_2+p_1) - (h_1+p_1)\Phi(s_1^*)] \cdot \phi(s_2) + (h_1+p_1)\int_{s_2}^{D} \phi(s_1^* + s_2 - u)\phi(u)du \\
&\geq (h_1+p_1)(D-s_2)\gamma^2 \\
&\geq \frac{\gamma^2(h_1+p_1)}{\Gamma(h_2+p_1)} \triangleq \sigma_2''.
\end{aligned}$$

Therefore, according to Lemma B.5, we know that with probability at least $1-\delta$,

$$|\bar{s}_{L,2} - s_2^*| \leq \mathcal{O}\left(\sqrt{\frac{\sqrt{L\log(TD/\delta)}}{L}} + \sqrt{\frac{\log(1/\delta)}{L}}\right) = \mathcal{O}\left(L^{-\frac{1}{4}}\log^{\frac{1}{4}}(TD/\delta)\right), \tag{16}$$

which finishes the proof.

□

Now we are ready to prove our main result Theorem 3.2 in the central planner setting. For completeness, we restate the theorem as follows.

Theorem A.7 (Restatement of Theorem 3.2)**.** *Algorithm 1 guarantees that with probability at least $1-2\delta$, the strategy converges to the optimal base-stock policy with the following rate:*

$$\begin{aligned}
|s_{M,1} - s_1^*| &\leq \mathcal{O}\left(\sqrt{\log(T/\delta)/T}\right), \\
|s_{M,2} - s_2^*| &\leq \mathcal{O}\left(T^{-1/4}\log^{1/4}(T/\delta)\right),
\end{aligned}$$

with $M = \mathcal{O}(\log T)$ the total number of epochs. Picking $\delta = 1/T^2$, Algorithm 1 also guarantees that $\mathbb{E}[\mathrm{Reg}_T] \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

Proof. We first prove the convergence of $s_{M,1}$ and $s_{M,2}$. According to Equation (11) and Lemma A.6, we know that with probability at least $1-2\delta$, for all $m \in [M]$,

$$|s_{m,1} - s_1^*| \leq \frac{C_1}{\Gamma}\sqrt{\frac{\log(TD/\delta)}{L_{m-1}}} = \mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{2^m}}\right), \tag{17}$$

$$|s_{m,2} - s_2^*| \leq \mathcal{O}\left(L_{m-1}^{-\frac{1}{4}}\log^{\frac{1}{4}}(T/\delta)\right) = \mathcal{O}\left(2^{-\frac{m}{4}}\log^{\frac{1}{4}}(T/\delta)\right). \tag{18}$$

Picking $m = M$ and noticing the fact that $2^m = \Theta(T)$ prove the convergence of $s_{M,1}$ and $s_{M,2}$.

According to Equation (15), picking $\delta = \frac{1}{T^2}$, we obtain that

$$L_m \cdot \mathbb{E}\left[H(s_1^*, s_{m,2}) - H(s_1^*, s_2^*)\right]$$

Algorithm 4 Online Newton Step

Input: learning rate $\eta > 0$, perturbation $\varepsilon > 0$.
Initialize: $x_1 = x_0$ arbitrarily.
for $t = 1$ *to* T **do**
  Choose action $x_t \in \mathcal{X}$ and observe $g_t = \nabla f_t(x_t)$.
  Update $x_{t+1} = \Pi_{\mathcal{X}}^{M_t}(x_t - \eta M_t^{-1} g_t)$, where $M_t = \sum_{s=1}^{t} g_s g_s^\top + \varepsilon I$.
end

$$= \mathbb{E}\left[\sum_{t\in I_m} H(s_1^*, s_{m,2}) - \sum_{t\in I_m} H(s_1^*, s_2^*)\right] \le \mathcal{O}\left(\sqrt{L_m \log T}\right).$$

Furthermore, according to the $\max\{h_1, p_1\}$-Lipschitzness of $H(\cdot, s_2)$ for any s_2 and Equation (17), we know that

$$\begin{aligned} & L_m \cdot \mathbb{E}\left[H(s_{m,1}, s_{m,2}) - H(s_1^*, s_2^*)\right] \\ & \le L_m \cdot \mathbb{E}\left[H(s_1^*, s_{m,2}) - H(s_1^*, s_2^*)\right] + L_m \mathcal{O}\left(|s_{m,1} - s_1^*|\right) \\ & \le \mathcal{O}(\sqrt{L_m \log T}). \end{aligned} \tag{19}$$

To further show that the expected regret is also well-bounded, as proven in Lemma 3.1, within each epoch, there is only constant number of rounds such that $\widetilde{H}_t \neq \widehat{H}_t(s_{m,1}, s_{m,2})$, $t \in I_m$. Therefore, picking $\delta = 1/T^2$, we have,

$$\begin{aligned} \mathbb{E}\left[\mathrm{Reg}_T\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{H}_t - \sum_{t=1}^{T} H(s_1^*, s_2^*)\right] \\ &= \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(\widetilde{H}_t - H(s_{m,1}, s_{m,2})\right)\right] + \mathbb{E}\left[\sum_{m=1}^{M}\left(\sum_{t\in I_m} H(s_{m,1}, s_{m,2}) - \sum_{t\in I_m} H(s_1^*, s_2^*)\right)\right] \\ &= \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(\widetilde{H}_t - \widehat{H}_t(s_{m,1}, s_{m,2})\right)\right] + \mathbb{E}\left[\sum_{m=1}^{M}\left(\sum_{t\in I_m} H(s_{m,1}, s_{m,2}) - \sum_{t\in I_m} H(s_1^*, s_2^*)\right)\right] \\ &\le \mathcal{O}(\log T) + \mathbb{E}\left[\sum_{m=1}^{M} \mathcal{O}\left(\sqrt{L_m \log(TD/\delta)}\right)\right] \qquad \text{(Equation (19))} \\ &\le \mathcal{O}(\sqrt{T \log T}), \end{aligned}$$

where the last inequality is because of the exponential length scheduling of the epochs. □

B Omitted proofs in Section 3.2

B.1 ONS and Lazy ONS algorithm

We show full pseudo code of the classic ONS algorithm in Algorithm 4 and our proposed lazy ONS algorithm in Algorithm 5.

B.2 $\widehat{H}_{t,2}^{\mathbf{c},\omega}(x)$ satisfies Property 1

Lemma B.1. *Suppose that demand distribution $\mathcal{D}$ satisfies and Assumption 2. The stochastic function $\widehat{H}_{t,2}^{c,\omega}$ defined in Equation (8) satisfies Property 1 with $B = \frac{\max\{\omega^2, h_2^2\}}{\gamma(h_2+\omega)}$ for all $t \in [T]$.*

Proof. Direct calculation shows that

$$H_2^{c,\omega}(x) \triangleq \mathbb{E}\left[\widehat{H}_{t,2}^{c,\omega}(x)\right] = (h_2+\omega)\mu + (h_2+\omega)\int_0^x \left(\Phi(u) - \frac{\omega}{h_2+\omega}\right) du,$$

Algorithm 5 Online Newton Step with lazy update

Input: learning rate η, perturbation $\varepsilon > 0$, total horizon T.
Initialize: $\widehat{x}_1 = x_1 = x_0$ arbitrarily, $k = 0$.
for $t = 1$ *to* T **do**
  if $t = 2^k$ **then**
    $k \leftarrow k + 1$
    $\widehat{x}_k = \frac{1}{t}\sum_{s=1}^{t} x_t \in \mathcal{X}$.
  end
  Choose action $w_t = \widehat{x}_k \in \mathcal{X}$ and observe f_t.
  Set $g_t = \nabla f_t(x_t)$
  Update $x_{t+1} = \Pi_{\mathcal{X}}^{M_t}(x_t - \eta M_t^{-1} g_t)$, where $M_t = \sum_{s=1}^{t} g_s g_s^\top + \varepsilon I$.
end

where $\mu = \mathbb{E}_{d' \sim \mathcal{D}}[d']$. Also it is direct to see that the minimizer of $H_2^{\mathrm{c},\omega}(x)$ is $x^* = \Phi^{-1}\left(\frac{\omega}{\omega+h_2}\right)$. Taking the gradient of $\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)$, we have:

$$\nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x) = (h_1 + p_1)\mathbb{I}\{x \geq d_t\} - p_1,$$
$$\nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x) \cdot \nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x) = (h_1 + p_1)^2 \mathbb{I}\{x \geq d_t\} - 2p_1(h_1 + p_1)\mathbb{I}\{x \geq d_t\} + p_1^2.$$

Taking expectation of the above two equations, we have

$$\mathbb{E}\left[\nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x)\right] = (h_1 + p_1)\Phi(x) - p_1,$$
$$\mathbb{E}\left[\nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x) \cdot \nabla \widehat{H}_{t,2}^{\mathrm{c},\omega}(x)\right] = p_1^2 + \Phi(x)(h_1^2 - p_1^2).$$

To show that $\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)$ satisfies Property 1, we first consider the case $x \geq x^* = \Phi^{-1}\left(\frac{\omega}{\omega+h_1}\right)$. In this case, we need to find $B > 0$ such that for all $x \geq x^*$:

$$B \geq \frac{(x - x^*)(\omega^2 + \Phi(x)(h_1^2 - \omega^2))}{(\omega + h_1)\Phi(x) - \omega}.$$

Using Assumption 2, we have $(x - x^*) \leq \frac{1}{\gamma(h_1+\omega)}((h_1 + \omega)\Phi(x) - \omega)$, which means that

$$\frac{(x - x^*)(\omega^2 + \Phi(x)(h_1^2 - \omega^2))}{(\omega + h_1)\Phi(x) - \omega} \leq \frac{\omega^2 + \Phi(x)(h_1^2 - \omega^2)}{\gamma(h_1 + \omega)} \leq \frac{\max(\omega^2, h_1^2)}{\gamma(h_1 + \omega)}.$$

Choosing $B \geq \frac{\max(\omega^2, h_1^2)}{\gamma(h_1+\omega)}$ satisfies Property 1. The second case where $x \leq x^*$ can be proved in a similar way. Therefore, we show that $\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)$ satisfies Property 1. □

As claimed in Footnote 2, we show in the following lemma that even when the demand distribution is discrete and the expected loss function is *not* strongly convex, the realized loss function $\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)$ also satisfies Property 1.

Lemma B.2. *Suppose that demand distribution is supported on finite values $d_i > 0$ with probability $w_i > 0$, $i \in [k]$, $\sum_{i=1}^{k} w_i = 1$ and $d_1 < d_2 < \ldots < d_k$. Also suppose that there exists a unique $i^* \in [k]$ such that $\Phi(d_{i^*-1}) < \frac{\omega}{h_1+\omega}$ and $\Phi(d_{i^*}) > \frac{\omega}{h_1+\omega}$. Let $\theta = \min\{\Phi(d_{i^*}) - \frac{\omega}{h_1+\omega}, \frac{\omega}{h_1+\omega} - \Phi(d_{i^*-1})\}$. The stochastic function $\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)$ defined in Equation (8) satisfies Property 1 with $B = \frac{\max_{i\in[k]} d_i \cdot \max\{\omega^2, h_1^2\}}{\theta(h_1+\omega)}$.*

Proof. We first show that $\mathbb{E}_{d_t \sim \mathcal{D}}[\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)]$ is not strongly convex. In fact, direct calculation shows that

$$\mathbb{E}_{d_t \sim \mathcal{D}}\left[\widehat{H}_{t,2}^{\mathrm{c},\omega}(x)\right] = \sum_{i=1}^{k} w_i \left(\omega(x - d_i)^+ + h_1(x - d_i)^-\right),$$

which is a piece-wise linear function, thus not strongly convex.

To show that $\widehat{H}^{c,\omega}_{t,2}(x)$ satisfies Property 1, direct calculation shows that

$$\mathbb{E}\left[\nabla \widehat{H}^{c,\omega}_{t,1}(x)\right] = (h_1+\omega)\Phi(x) - \omega,$$
$$\mathbb{E}\left[\nabla \widehat{H}^{c,\omega}_{t,2}(x) \cdot \nabla \widehat{H}^{c,\omega}_{t,2}(x)\right] = \omega^2 + \Phi(x)(h_1^2 - \omega^2).$$

It is also direct to see that the minimizer of $\mathbb{E}[\widehat{H}^{c,\omega}_{t,2}(x)]$ is $x^* = d_{i^*}$. When $x \geq x^*$, we need to show that there exists $B > 0$ such that for all $x > x^*$,

$$B \geq \frac{(x-x^*)(\omega^2 + \Phi(x)(h_1^2-\omega^2))}{(\omega+h_1)\Phi(x) - \omega}. \tag{20}$$

Note that for $x \geq x^*$, $\Phi(x) \geq \Phi(x^*) \geq \theta + \frac{\omega}{\omega+h_1}$ and $x - x^* \leq \max_{i\in[k]} d_i$, therefore, we have

$$\frac{(x-x^*)(\omega^2 + \Phi(x)(h_1^2-\omega^2))}{(\omega+h_1)\Phi(x) - \omega} \leq \frac{\max_{i\in[k]} d_i \cdot \max\{\omega^2, h_1^2\}}{\theta(\omega+h_1)},$$

meaning that $B = \frac{\max_{i\in[k]} d_i \cdot \max\{\omega^2,h_1^2\}}{\theta(\omega+h_1)}$ satisfies Equation (20). Similarly, when $x \leq x^*$, we can also show that $B = \frac{\max_{i\in[k]} d_i \cdot \max\{\omega^2,h_1^2\}}{\theta(\omega+h_1)}$ satisfies that

$$B \geq \frac{(x^*-x)(\omega^2 + \Phi(x)(h_1^2-\omega^2))}{\omega - (\omega+h_1)\Phi(x)}.$$

Combining both cases shows that $\widehat{H}^{c,\omega}_{t,2}(x)$ satisfies Property 1. □

B.3 ONS achieves $\mathcal{O}(\log T)$ regret when Property 1 is satisfied

Theorem B.3. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set with bounded diameter $\max_{x,x'\in\mathcal{X}} \|x - x'\| \leq J$. If $\{f_t\}_{t=1}^T$ satisfy Property 1 for some $B > 0$, $f_t : \mathcal{X} \mapsto \mathbb{R}$ and $\max \|\nabla f_t(x)\| \leq G$, Algorithm 4 with $\eta \geq 2B$ and $\varepsilon = 1/T$ ensures: $\mathbb{E}[\sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x^*)] \leq \mathcal{O}(d\log(GT) + J^2/2BT)$, where $x^* = \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}[f_t(x)]$.*

Proof. The first part of the proof follows the classic ONS proof: let $y_{t+1} = x_t - \eta M_t^{-1} g_t$ and we know that

$$y_{t+1} - x^* = x_t - x^* - \eta M_t^{-1} g_t,$$
$$M_t(y_{t+1} - x^*) = M_t(x_t - x^*) - \eta g_t.$$

Therefore, by definition of x_{t+1}, we know that

$$\|x_{t+1} - x^*\|^2_{M_t} \leq \|y_{t+1} - x^*\|^2_{M_t} = \|x_t - x^*\|^2_{M_t} - 2\eta\langle x_t - x^*, g_t\rangle + \eta^2 \|g_t\|^2_{M_t^{-1}},$$

where $\|x\|^2_M \triangleq x^\top M x$. Rearranging the terms, we know that

$$\langle x_t - x^*, g_t\rangle \leq \frac{\|x_t - x^*\|^2_{M_t} - \|x_{t+1} - x^*\|^2_{M_t}}{\eta} + \eta \|g_t\|^2_{M_t^{-1}}.$$

Taking summation over $t \in [T]$ using the definition of M_t, we know that

$$\sum_{t=1}^T \langle x_t - x^*, g_t\rangle \leq \frac{\|x_1 - x^*\|^2_{M_0}}{\eta} + \frac{1}{\eta}\sum_{t=1}^T (x_t - x^*)^\top g_t g_t^\top (x_t - x^*) + \eta \sum_{t=1}^T \|g_t\|^2_{M_t^*}.$$

By choosing $\varepsilon = \frac{1}{T}$, we have the first term bounded by $\mathcal{O}(\frac{J^2}{\eta T})$. For the third term, according to the assumption that $\|g_t\|_2 \leq G$ and Lemma 6 in (Hazan et al., 2007), we obtain that

$$\sum_{t=1}^T \|g_t\|^2_{M_t^*} \leq \log\left(\frac{\det(\sum_{t=1}^T g_t g_t^\top + \varepsilon I)}{\det(\varepsilon I)}\right) \leq d\log\left(\frac{G^2 T}{\varepsilon} + 1\right) \leq 4d\log(GT).$$

Finally, we consider the second term. Let $f(x) = \mathbb{E}_{f_t \sim \mathcal{F}}[f_t(x)]$. According to the convexity of f, we have for any $x, y \in \mathcal{X}$,

$$f(y) \geq f(x) + (y - x)^\top \nabla f(x).$$

Choosing $y = x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$ and using Property 1, we have

$$\begin{aligned} f(x^*) &\geq f(x) + (x^* - x)^\top \nabla f(x) \\ &\geq f(x) + 2(x^* - x)^\top \nabla f(x) + \frac{1}{B}(x - x^*)^\top \mathbb{E}_{f_t \sim \mathcal{F}} \left[\nabla f_t(x) \nabla f_t(x)^\top\right] (x - x^*). \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^T f_t(x_t)\right] - \mathbb{E}\left[\sum_{t=1}^T f_t(x^*)\right] \\ &\leq 2\mathbb{E}\left[\sum_{t=1}^T \langle x_t - x^*, g_t \rangle\right] - \mathbb{E}\left[\sum_{t=1}^T \frac{1}{B}(x_t - x^*)^\top g_t g_t^\top (x_t - x^*)\right] \\ &\leq \mathcal{O}\left(\frac{J^2}{\eta T} + d \log(GT)\right) + \left(\frac{2}{\eta} - \frac{1}{B}\right) \mathbb{E}\left[\sum_{t=1}^T (x_t - x^*)^\top g_t g_t^\top (x_t - x^*)\right]. \end{aligned}$$

Choosing $\eta \geq 2B$ leads to the bound. □

B.4 Proof of Theorem B.4

Finally, we prove that Algorithm 5, a lazy version of Algorithm 4 which only updates the decisions $\mathcal{O}(\log T)$ times over T rounds, achieves $\mathcal{O}(\log^2 T)$ expected regret guarantee. This algorithm shares the same spirit of Algorithm 3 in (Sherman and Koren, 2021). We highlight again that the low switching property is important to achieve $\mathcal{O}(\log^2 T)$ regret bound in our two-echelon inventory control problem.

Theorem B.4. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set with bounded diameter $\max_{x,x' \in \mathcal{X}} \|x - x'\| \leq J$. If $\{f_t\}_{t=1}^T$ satisfy Property 1 for some $B > 0$, $f_t : \mathcal{X} \mapsto \mathbb{R}$ and $\max_{t \in [T], x \in \mathcal{X}} \|\nabla f_t(x)\| \leq G$, then Algorithm 5 with $\eta \geq 2B$, $\varepsilon = \frac{1}{T}$ guarantees that $\mathbb{E}[\sum_{t=1}^T f_t(w_t)] - \mathbb{E}[\sum_{t=1}^T f_t(x^*)] \leq \mathcal{O}(\log^2 T + \log G + J^2/B)$, where $x^* = \operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}_{f \sim \mathcal{F}}[f(x)]$. Moreover, the decision sequence $\{w_t\}_{t=1}^T$ only switches $\mathcal{O}(\log T)$ times.*

Proof. In Theorem B.3, we know that for any $t \in [T]$, the decision sequence $\{x_s\}_{s=1}^t$ generated by ONS (Algorithm 4) guarantees that,

$$\mathbb{E}\left[\sum_{s=1}^t f_s(x_s)\right] - \mathbb{E}\left[\sum_{s=1}^t f_s(x^*)\right] \leq \mathcal{O}(d \log GT + J^2/2BT),$$

where $x^* = \operatorname{argmin}_{f \sim \mathcal{F}} f(x)$. For any fixed t, using the convexity and stochasticity of f_t, we know that

$$\mathbb{E}\left[f_t(\widehat{x}_k) - f_t(x^*)\right] \leq \frac{1}{2^k} \sum_{s=1}^{2^k} \mathbb{E}\left[f_t(x_s) - f_t(x^*)\right] = \frac{1}{2^k} \sum_{s=1}^{2^k} \mathbb{E}\left[f_s(x_s) - f_s(x^*)\right] \leq \mathcal{O}\left(\frac{1}{2^k} \cdot \left(d \log(G 2^k) + \frac{J^2}{2^{k+1} B}\right)\right).$$

Taking a summation over all $t \in [T]$, we have

$$\mathbb{E}\left[\sum_{t=1}^T f_t(w_t)\right] - \mathbb{E}\left[\sum_{t=1}^T f_t(x^*)\right] = \mathbb{E}\left[\sum_{k=0}^{\log_2 T} \sum_{t \in I_k} (f_t(\widehat{w}_k) - f_t(x^*))\right] \leq \mathcal{O}\left(\log^2 T + \log G + \frac{J^2}{B}\right),$$

where I_k is the set of time index in the k-th epoch $[2^{k-1}, 2^k - 1]$. □

Lemma B.5. *Suppose that $\{f_t\}_{t=1}^T$ is a sequence of i.i.d convex functions drawn from a distribution $\mathcal{F}$ and each $f_t : \mathcal{X} \mapsto \mathbb{R}$ has the same bounded feasible domain $\max_{x,x' \in \mathcal{X}} |x - x'| \leq J$. Let $x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$ where $f(x) = \mathbb{E}_{f_t \sim \mathcal{F}}[f_t(x)]$*

Algorithm 6 Contract maker

Input: A set of realized demand value $S = \{d_1, \ldots, d_L\}$, learning rate η.
Construct empirical cumulative density function $\widehat{\Phi}_L(\cdot)$ using S.
Let s_L be the output of Algorithm 2 with input $\mathcal{D}$, $\widehat{\Phi}_L(\cdot)$ and η.
return $\omega_L = \frac{h_2 \widehat{\Phi}_L(s_L)}{1-\widehat{\Phi}(s_L)}$.

and suppose that $f(x)$ is σ-strongly convex. Suppose that $\{x_t\}_{t=1}^T$ be the decision sequence Algorithm 4 generates when the loss function sequence is $\{f_t\}_{t=1}^T$. Suppose that $\mathbb{E}\left[\sum_{t=1}^T f(x_t) - f(x^*)\right] \leq R$. Let $\bar{x}_1 = \frac{2}{T}\sum_{t=1}^{T/2} x_t$ and $\bar{x}_2 = \frac{1}{T}\sum_{t=1}^{T} x_t$. Then, with probability at least $1-\delta$, we have

$$|\bar{x}_1 - x^*| \leq \mathcal{O}\left(\sqrt{\frac{R}{T\sigma}} + J\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right),$$
$$|\bar{x}_2 - x^*| \leq \mathcal{O}\left(\sqrt{\frac{R}{T\sigma}} + J\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right).$$

Proof. According to the strong convexity of $f(x)$, we know that

$$\frac{\sigma}{2}\mathbb{E}\left[\sum_{t=1}^{T/2}|x_t - x^*|^2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T/2}(f(x_t) - f(x^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}(f(x_t) - f(x^*))\right] \leq R.$$

By Cauchy-Schwarz inequality and the fact that $\mathbb{E}[x^2] \geq \mathbb{E}[x]^2$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T/2}|x_t - x^*|\right] \leq \mathcal{O}\left(\sqrt{\frac{RT}{\sigma}}\right).$$

According to Azuma's inequality and the boundedness of x_t, x^*, we have with probability at least $1 - \frac{\delta}{2}$,

$$\sum_{t=1}^{T/2}|x_t - x^*| - \mathbb{E}\left[\sum_{t=1}^{T/2}|x_t - x^*|\right] \leq \mathcal{O}\left(J\sqrt{T\log\frac{1}{\delta}}\right). \tag{21}$$

Note that $\bar{x}_1 = \frac{2}{T}\sum_{t=1}^{T/2} x_t$. Therefore, with probability at least $1-\delta$,

$$\frac{T}{2}|\bar{x}_1 - x^*| \leq \sum_{t=1}^{T/2}|x_t - x^*| \leq \mathcal{O}\left(J\sqrt{T\log\frac{1}{\delta}} + \sqrt{\frac{RT}{\sigma}}\right).$$

Applying a similar analysis on $\bar{x}_2$ and a union bound finishes the proof. □

B.5 Algorithm for the contract maker

We show the pseudo code of the algorithm for the contract maker in Algorithm 6.

B.6 Proof of Lemma 3.3

Proof. According to Lemma A.6, we know that with probability at least $1-\delta$, $|\bar{s}_{L,2} - s_2^*| \leq \widetilde{\mathcal{O}}(L^{-\frac{1}{4}})$, where $\bar{s}_{L,2}$ is the output of Algorithm 2. To bound the difference between the contract coefficient ω returned by the third party and the optimal ω^*, note that $\omega = \frac{h_2\widehat{\Phi}_L(\bar{s}_{L,2})}{1-\widehat{\Phi}_L(\bar{s}_{L,2})}$. According to Lemma C.1, with probability at least $1-\delta$, it holds that

$$\left|\widehat{\Phi}_L(\bar{s}_{L,2}) - \Phi(s_2^*)\right| \leq \left|\widehat{\Phi}_L(\bar{s}_{L,2}) - \Phi_L(\bar{s}_{L,2})\right| + |\Phi(\bar{s}_{L,2}) - \Phi(s_2^*)|$$

$$\leq \sqrt{\frac{1}{2L}\log\frac{2}{\delta}} + \Gamma\left|\bar{s}_{L,2} - s_2^*\right|$$
$$\leq C_3 \log(T/\delta) L^{-\frac{1}{4}}, \tag{22}$$

where the last inequality is due to Lemma A.6 and $C_3 > 0$ is a universal constant. Moreover, note that according to Equation (14), we know that $\Phi(s_2^*) \leq \frac{p_1}{h_2+p_1}$, meaning that $\frac{1}{1-\Phi(s_2^*)} \leq \frac{h_2+p_1}{h_2}$. Combining with Equation (22), we know that with probability $1-\delta$

$$1 - \widehat{\Phi}_L(\bar{s}_{L,2}) \geq 1 - \Phi(s_2^*) - C_3\log(T/\delta)L^{-\frac{1}{4}} \geq \frac{h_2}{h_2+p_1} - C_3\log(T/\delta)L^{-\frac{1}{4}} \geq \frac{h_2}{2(h_2+p_1)},$$

where the last inequality holds when $L \geq C_4 \triangleq \frac{16(h_2+p_1)^4 C_3^4 \log^4(T/\delta)}{h_2^4}$. Also it holds that $\omega^* = \frac{h_2\Phi(s_2^*)}{1-\Phi(s_2^*)} \leq \frac{h_2}{1-\frac{p_1}{h_2+p_1}} = h_2 + p_1$.

Therefore, we obtain that

$$|\omega - \omega^*| \leq \left|\frac{h_2\widehat{\Phi}_L(\bar{s}_{L,2})}{1-\widehat{\Phi}_L(\bar{s}_{L,2})} - \frac{h_2\Phi(s_2^*)}{1-\Phi(s_2^*)}\right| \leq h_2\left|\frac{\widehat{\Phi}_L(\bar{s}_{L,2}) - \Phi(s_2^*)}{(1-\widehat{\Phi}_L(\bar{s}_{L,2}))(1-\Phi(s_2^*))}\right| \leq \mathcal{O}(L^{-\frac{1}{4}}\log(T/\delta)),$$

which further shows that $\omega \in [0, \omega^* + \mathcal{O}(L^{-\frac{1}{4}}\log(T/\delta))] \subseteq [0, h_2 + p_1 + \mathcal{O}(L^{-\frac{1}{4}}\log(T/\delta))]$. □

B.7 Proof of Theorem 3.4

Proof. We first show the convergence on the inventory level decisions of Agent 2. We consider each epoch I_m separately. As Agent 1 keeps his desired inventory level within each epoch, according to Lemma 3.1, there are only constant number of rounds at the beginning of epoch I_m such that $o_t \neq d_t$. With a slight abuse of notation, define

$$\widehat{H}_{t,2}^{\mathrm{c}}(s_2) = h_2(s_2 - d_t)^+ + \omega_m(s_2 - d_t)^-,$$

for $t \in I_m$. According to Lemma B.1, we know that $\widehat{H}_{t,2}^{\mathrm{c}}$ satisfies Property 1 with a specific choice of $B > 0$. Therefore, according to Algorithm 3 and Lemma B.1, Agent 2 is using Algorithm 5 within each epoch with respect to $\widehat{H}_{t,2}^{\mathrm{c}}$ except for constant number of rounds at the beginning of the epoch. According to Theorem B.4 and Lemma 3.1, we know that the expected regret of Agent 2 within epoch I_m is bounded as follows: picking $\delta = \frac{1}{T^2}$, for any $s_2 \in [d, D]$,

$$\mathbb{E}\left[\sum_{t\in I_m}\left(\widetilde{H}_{t,2}^{\mathrm{c}} - \mathring{H}_{t,2}^{\mathrm{c}}(s_2)\right)\right] \leq \mathcal{O}\left(\log^2 T + \log T\right) + \mathcal{O}(1) = \mathcal{O}(\log^2 T). \tag{23}$$

As the total regret is upper bounded by the sum of the regrets in each epoch $m \in [M]$, $M = \mathcal{O}(\log T)$, we know that

$$\mathbb{E}\left[\mathrm{Reg}_{T,2}\right] = \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(\widetilde{H}_{t,2}^{\mathrm{c}} - \mathring{H}_{t,2}^{\mathrm{c}}(\mathring{s}_2^*)\right)\right] \leq \mathcal{O}(\log^3 T).$$

Next, we consider the convergence of $s_{T,2}$ and bound the term $|s_{T,2} - s_2^*|$. More generally, let e_m be the last round of epoch m and we bound $|s_{e_m,2} - s_2^*|$. First, according to the analysis in Lemma 3.3 with a union bound, we know that if ω_m is generated by Algorithm 2, with probability at least $1-\delta$, for any epoch index $m \in [M]$,

$$|\omega_m - \omega^*| \leq \mathcal{O}\left(L_m^{-\frac{1}{4}}\log\frac{T}{\delta}\right). \tag{24}$$

In addition, note that according to the dynamic of Agent 1 and Agent 2, there are $\Theta(2^m \cdot L_1)$ rounds in the epoch I_m where Agent 1 keeps choosing her inventory level to be $s_{m,1}$, Agent 2 keeps choosing her inventory level to be $s_{e_m,2}$ and the contract coefficient is ω_m. Define the set of these rounds to be $\mathcal{T}_m$. In addition, define the expected loss function $H_{t,2}^{\mathrm{c}}(s_2) = h_2\mathbb{E}_{x\sim\mathcal{D}}\left[(s_2 - x)^+\right] + \omega_m\mathbb{E}_{x\sim\mathcal{D}}\left[(s_2 - x)^-\right]$ for $t \in \mathcal{T}_m$ and $\bar{s}_{m,2}^* = \mathrm{argmin}_{s_2} H_{t,2}^{\mathrm{c}}(s_2) = \Phi^{-1}\left(\frac{\omega_m}{\omega_m+h_2}\right)$,

where the second equality is by direct calculation. In addition, recall that $\widehat{H}^{\mathrm{c}}_{t,2}(s_2) = h_2(s_2 - d_t)^+ + \omega_m(s_2 - d_t)^-$. Then by choosing $\delta = \frac{1}{T^2}$, we know that for all $m \in [M]$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{T}_m} H^{\mathrm{c}}_{t,2}(s_{e_m,2}) - \sum_{t\in\mathcal{T}_m} H^{\mathrm{c}}_{t,2}(\bar{s}^*_{m,2})\right] \\
&\le \mathbb{E}\left[\sum_{t\in I_m} H^{\mathrm{c}}_{t,2}(s_{t,2}) - \sum_{t\in I_m} H^{\mathrm{c}}_{t,2}(\bar{s}^*_{m,2})\right] \\
&= \mathbb{E}\left[\sum_{t\in I_m} \widehat{H}^{\mathrm{c}}_{t,2}(s_{t,2}) - \sum_{t\in I_m} \widehat{H}^{\mathrm{c}}_{t,2}(\bar{s}^*_{m,2})\right] \\
&\le \mathcal{O}(\log^2 T). \qquad \text{(Lemma 3.1 and Theorem B.4)}
\end{aligned} \tag{25}$$

In addition, according to Lemma C.4, we know that $H^{\mathrm{c}}_{t,2}(s_2)$ is strongly convex in s_2 with parameter $\sigma_m = \gamma(h_2 + \omega_m)$. Therefore, according to Lemma B.5, we have with probability at least $1-\delta$, for all $m \in [M]$,

$$\left|s_{e_m,2} - \bar{s}^*_{m,2}\right| \le \mathcal{O}\left(L_m^{-\frac{1}{2}} \log(T/\delta)\right) \tag{26}$$

Now we are ready to bound $|s_{e_m,2} - s^*_2|$. Recall that $s^*_2 = \Phi^{-1}(\omega^*/(\omega^* + h_2))$. Therefore, with probability at least $1 - 2\delta$, for all $m \in [M]$,

$$\begin{aligned}
|s_{e_m,2} - s^*_2| &\le |s_{e_m,2} - \bar{s}^*_{m,2}| + |\bar{s}^*_{m,2} - s^*_2| \\
&\le \mathcal{O}(L_m^{-\frac{1}{2}} \log(T/\delta)) + \left|\Phi^{-1}\left(\frac{\omega_m}{\omega_m + h_2}\right) - \Phi^{-1}\left(\frac{\omega^*}{\omega^* + h_2}\right)\right| && \text{(Equation (26))} \\
&\le \mathcal{O}(L_m^{-\frac{1}{2}} \log(T/\delta)) + \frac{1}{\gamma}\left|\frac{\omega_m}{\omega_m + h_2} - \frac{\omega^*}{\omega^* + h_2}\right| && \text{(Assumption 2)} \\
&\le \mathcal{O}(L_m^{-\frac{1}{2}} \log(T/\delta)) + \mathcal{O}(|\omega_m - \omega^*|) \\
&\le \mathcal{O}(L_m^{-\frac{1}{4}} \log(T/\delta)),
\end{aligned} \tag{27}$$

where the last inequality is due to Equation (24). Applying $m = M$ shows that $|s_{e_M,2} - s^*_2| = |s_{T,2} - s^*_2| \le \mathcal{O}(T^{-\frac{1}{4}} \log(T/\delta))$, which finishes the proof for the convergence of Agent 2.

In addition, according to Equation (25) and Cauchy-Schwarz inequality, we know that within epoch I_m,

$$\frac{\sigma_m}{2}\mathbb{E}\left[\sum_{t\in I_m} |s_{t,2} - \bar{s}^*_{m,2}|^2\right] \le \mathcal{O}(\log^2 T) \Rightarrow \mathbb{E}\left[\sum_{t\in I_m} |s_{t,2} - \bar{s}^*_{m,2}|\right] \le \mathcal{O}(\sqrt{L_m}\log T). \tag{28}$$

In addition, based on the boundedness of $s_{t,2}$ and $\bar{s}^*_{m,2}$, according to Hoeffding-Azuma's inequality, similar to Equation (21), with probability at least $1-\delta$, we know that for all $m \in [M]$,

$$\sum_{t\in I_m} |s_{t,2} - \bar{s}^*_{m,2}| - \mathbb{E}\left[\sum_{t\in I_m} |s_{t,2} - \bar{s}^*_{m,2}|\right] \le \mathcal{O}\left(\sqrt{L_m \log\frac{M}{\delta}}\right). \tag{29}$$

Therefore, combining Equation (29) and Equation (27), with probability at least $1-\delta$, for all $m \in [M]$,

$$\begin{aligned}
\sum_{t\in I_m} |s_{t,2} - s^*_2| &\le \sum_{t\in I_m} \left(|s_{t,2} - \bar{s}^*_{m,2}| + |\bar{s}^*_{m,2} - s^*_2|\right) \\
&\le \mathcal{O}\left(\sqrt{L_m \log\frac{M}{\delta}}\right) + \mathcal{O}\left(L_m^{\frac{3}{4}} \log(T/\delta)\right) = \mathcal{O}\left(L_m^{\frac{3}{4}} \log(T/\delta)\right).
\end{aligned} \tag{30}$$

For Agent 1, as $s_{m,1} = \widehat{\Phi}_{m-1}^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)$, using Lemma C.3, we know that with probability at least $1-\delta$, for any $m \in [M]$,

$$|s_{m,1} - s_1^*| = \left|\widehat{\Phi}_{m-1}^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right) - \Phi^{-1}\left(\frac{h_2+p_1}{h_1+p_1}\right)\right| \le C_0\sqrt{\frac{\log(2TD/\delta)}{L_{m-1}}} = \mathcal{O}\left(\sqrt{\frac{\log(2TD/\delta)}{2^m}}\right) \tag{31}$$

Again, setting $m = M$ proves the convergence of $s_{M,1}$.

Finally, we analyze the regret of Agent 1. For $t \in I_m$, define

$$\begin{aligned}
\widehat{H}_{t,1}^{\mathrm{c}}(s_1, s_2) &= h_1(\mathring{s}_{t,1} - d_t)^+ + p_1(\mathring{s}_{t,1} - d_t)^- - \omega_m(s_2 - d_t)^-, \\
H_{t,1}^{\mathrm{c}}(s_1, s_2) &= \mathbb{E}_{x\sim\mathcal{D}}\left[h_1(\widehat{s}_{t,1} - x)^+ + p_1(\widehat{s}_{t,1} - x)^-\right] - \omega_m\mathbb{E}_{x\sim\mathcal{D}}\left[(s_2 - x)^-\right],
\end{aligned}$$

where $\mathring{s}_{t,1} = s_1$ if $s_2 > d_{t-1}$ and $\mathring{s}_{t,1} = s_1 + s_2 - d_{t-1}$ otherwise. With the choice $\delta = \frac{1}{T^2}$, direct calculation shows that, for all s_1, $\widehat{H}_{t,1}^{\mathrm{c}}(s_1, \cdot)$ and $H_{t,1}^{\mathrm{c}}(s_1, \cdot)$ are $\mathcal{O}(\log T)$-Lipschitz according to Lemma 3.3. Based on Lemma 3.1, we know that within each epoch, except for constant number of rounds, Agent 1 can achieve her intended inventory level and $o_t = d_t$. Therefore, by choosing $\delta = \frac{1}{T^2}$, we know that

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Reg}_{T,1}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\widetilde{H}_{t,1}^{\mathrm{c}}\right] - \mathbb{E}\left[\sum_{t=1}^{T}\mathring{H}_{t,1}^{\mathrm{c}}(\mathring{s}_1^*)\right] \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\widehat{H}_{t,1}^{\mathrm{c}}(s_{m,1}, s_{t,2})\right] - \min_{s_1}\mathbb{E}\left[\sum_{t=1}^{T}\widehat{H}_{t,1}^{\mathrm{c}}(s_1, s_{t,2})\right] + \widetilde{\mathcal{O}}(1) && \text{(Lemma 3.1)} \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\widehat{H}_{t,1}^{\mathrm{c}}(s_{m,1}, s_2^*)\right] - \min_{s_1}\mathbb{E}\left[\sum_{t=1}^{T}\widehat{H}_{t,1}^{\mathrm{c}}(s_1, s_2^*)\right] + \mathbb{E}\left[\sum_{m}\sum_{t\in I_m}\widetilde{\mathcal{O}}(|s_{t,2} - s_2^*|)\right] + \widetilde{\mathcal{O}}(1) \\
& && \text{(Lipschitzness of } \widehat{H}_{t,1}^{\mathrm{c}}(s_1, \cdot)) \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}H_{t,1}^{\mathrm{c}}(s_{m,1}, s_2^*)\right] - \min_{s_1}\mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}H_{t,1}^{\mathrm{c}}(s_1, s_2^*)\right] + \widetilde{\mathcal{O}}(T^{\frac{3}{4}}) && \text{(Equation (30))} \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}H_{t,1}^{\mathrm{c}}(s_{m,1}, s_2^*)\right] - \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}H_{t,1}^{\mathrm{c}}(s_1^*, s_2^*)\right] + \widetilde{\mathcal{O}}(T^{\frac{3}{4}}) && (s_1^* \text{ is the minimizer of } H_{t,1}^{\mathrm{c}}(s_1, s_2^*)) \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\widetilde{\mathcal{O}}\left(2^m \cdot |s_{m,1} - s_1^*|\right)\right] + \widetilde{\mathcal{O}}(T^{\frac{3}{4}}) && (\max\{h_1, p_1\}\text{-Lipschitzness of } H_{t,1}^{\mathrm{c}}(\cdot, s_2^*)) \\
&\le \mathbb{E}\left[\sum_{m=1}^{M}\widetilde{\mathcal{O}}\left(\sqrt{2^m}\right)\right] + \widetilde{\mathcal{O}}(T^{\frac{3}{4}}) && \text{(Equation (31))} \\
&\le \widetilde{\mathcal{O}}(T^{\frac{3}{4}}),
\end{aligned}$$

which finishes the proof.

□

B.8 Proof of Theorem 3.5

Proof. Note that from Equation (15) and the convexity of $H(s_1^*, s_2)$ in s_2, let $\bar{s}_{m,2}$ be the output of the contract maker Algorithm 2, we know that

$$\mathbb{E}\left[\sum_{t\in I_m}H(s_1^*, \bar{s}_{m,2}) - \sum_{t\in I_m}H(s_1^*, s_2^*)\right] \le \mathcal{O}\left(\sqrt{L_m \log T}\right). \tag{32}$$

Let $\bar{s}_{m,2}^* = \Phi^{-1}(\frac{\omega_m}{\omega_m+h_2})$. First, we bound $\sum_{t\in I_m}|s_{t,2} - \bar{s}_{m,2}^*|$ for each epoch m. Note that according to the analysis

in Equation (25), we know that

$$\mathbb{E}\left[\sum_{t\in I_m} H^{\mathrm{c}}_{t,2}(s_{t,2}) - \sum_{t\in I_m} H^{\mathrm{c}}_{t,2}(\bar{s}^*_{m,2})\right] \le \mathcal{O}(\log^2 T).$$

According to Equation (28) and Equation (29), we know that with probability $1-\delta$, for all $m \in [M]$,

$$\sum_{t\in I_m} \left|s_{t,2} - \bar{s}^*_{m,2}\right| \le \mathcal{O}\left(\sqrt{L_m}\log\frac{T}{\delta}\right). \tag{33}$$

Next, we bound $\sum_{t\in I_m}|\bar{s}^*_{m,2} - \bar{s}_{m,2}|$ for each epoch m. Define $\widetilde{s}_{m,2} = \widehat{\Phi}^{-1}_{m-1}(\frac{\omega_m}{\omega_m+h_2})$. Note that $\omega_m = \frac{h_2\widehat{\Phi}_{m-1}(\bar{s}_{m,2})}{1-\widehat{\Phi}_{m-1}(\bar{s}_{m,2})}$. Let $\{d_k\}_{k=1}^{L_{m-1}}$ be the demand samples realized in epoch I_{m-1} and let $\{d'_k\}_{k=1}^{L_{m-1}}$ be the sorted sequence in non-decreasing order. Then, with probability at least $1-\delta$, for each $m\in[M]$,

$$\begin{aligned}
&\sum_{t\in I_m}|\widetilde{s}_{m,2} - \bar{s}_{m,2}| = L_m\cdot\left|\widehat{\Phi}^{-1}_{m-1}\left(\widehat{\Phi}_{m-1}(\bar{s}_{m,2})\right) - \bar{s}_{m,2}\right| \\
&\le L_m\cdot\max_{k\in[L_{m-1}]}|d'_k - d'_{k-1}| \\
&\le L_m\cdot\frac{1}{\gamma}\max_{k\in[L_{m-1}]}\left|\Phi(d'_k) - \Phi(d'_{k-1})\right| \\
&\le \frac{2}{\gamma}\log\frac{L_{m-1}MT}{\delta},
\end{aligned} \tag{34}$$

where the last inequality is due to Equation (38).

Then, we bound the term $\sum_{t\in I_m}|\bar{s}^*_{m,2} - \widetilde{s}_{m,2}|$. According to Lemma C.3, with probability at least $1-\delta$, for each $m\in[M]$,

$$\sum_{t\in I_m}|\bar{s}^*_{m,2} - \widetilde{s}_{m,2}| = L_m\left|\Phi^{-1}\left(\frac{\omega_m}{\omega_m+h_2}\right) - \widehat{\Phi}^{-1}_{m-1}\left(\frac{\omega_m}{\omega_m+h_2}\right)\right| \le \widetilde{\mathcal{O}}\left(\frac{L_m}{\gamma}\sqrt{\frac{\log(1/\delta)}{L_{m-1}}}\right). \tag{35}$$

Therefore, according to the Lipschitzness of $H(s_1,s_2)$ in both parameters, picking $\delta = \frac{1}{T^3}$, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, s_{t,2}) - \sum_{t\in I_m} H(s^*_1, s^*_2)\right] \\
&= \mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, s_{t,2}) - \sum_{t\in I_m} H(s^*_1, \bar{s}^*_{m,2})\right] + \mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, \bar{s}^*_{m,2}) - \sum_{t\in I_m} H(s^*_1, \widetilde{s}_{m,2})\right] \\
&\quad + \mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, \widetilde{s}_{m,2}) - \sum_{t\in I_m} H(s^*_1, \bar{s}_{m,2})\right] + \mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, \bar{s}_{m,2}) - \sum_{t\in I_m} H(s^*_1, s^*_2)\right] \\
&\le \sum_{t\in I_m}\left(\widetilde{\mathcal{O}}\left(|s_{t,2} - \bar{s}^*_{m,2}|\right) + \widetilde{\mathcal{O}}\left(|\bar{s}^*_{m,2} - \widetilde{s}_{m,2}|\right) + \widetilde{\mathcal{O}}\left(|\widetilde{s}_{m,2} - \bar{s}_{m,2}|\right)\right) + \mathcal{O}(1) \\
&\quad + \mathbb{E}\left[\sum_{t\in I_m} H(s^*_1, \bar{s}_{m,2}) - \sum_{t\in I_m} H(s^*_1, s^*_2)\right] \\
&\le \widetilde{\mathcal{O}}(\sqrt{L_m}),
\end{aligned}$$

where the last inequality is by combining Equation (32), Equation (33), Equation (34) and Equation (35). Finally, note that from Equation (31), we know that for all $m\in[M]$, $\mathbb{E}[|s_{m,1} - s^*_1|] \le \widetilde{\mathcal{O}}(1/\sqrt{L_m})$. Again using the Lipschitzness of $H(s_1,s_2)$, we can obtain that for all $m\in[M]$,

$$\mathbb{E}\left[\sum_{t\in I_m} H(s_{m,1}, s_{t,2}) - \sum_{t\in I_m} H(s^*_1, s^*_2)\right] \le \widetilde{\mathcal{O}}(\sqrt{L_m}).$$

Taking summation over all $m \in [M]$, we know that

$$\mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(H(s_{m,1},s_{t,2})-H(s_1^*,s_2^*)\right)\right] \le \widetilde{\mathcal{O}}(\sqrt{T}).$$

Finally, according to Lemma 3.1, as both agents only changes their decision $\widetilde{\mathcal{O}}(1)$ number of rounds and within each epoch, we know that only constant number of round such that the desired inventory level can not be realized and $d_t \neq o_t$. Therefore, we can obtain that

$$\begin{aligned}\mathbb{E}\left[\mathrm{Reg}_T\right] &= \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(\widetilde{H}_t - H(s_1^*,s_2^*)\right)\right]\\ &\le \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t\in I_m}\left(H(s_{m,1},s_{t,2})-H(s_1^*,s_2^*)\right)\right] + \widetilde{\mathcal{O}}(1) \le \widetilde{\mathcal{O}}(\sqrt{T}),\end{aligned}$$

which finishes the proof. □

C Auxiliary lemmas

In this section, we introduce several lemmas that are useful in the analysis. The first three lemmas show the properties of the empirical density function and the true density function. Suppose in epoch I with $|I| = L$, we receive the demand $d_1, d_2, \ldots, d_L$. Define the empirical cumulative density function $\widehat{\Phi}_L(\cdot)$ constructed by $\{d_i\}_{i=1}^L$ as

$$\widehat{\Phi}_L(x) = \frac{1}{L}\sum_{i=1}^{L}\mathbb{I}\{d_i \le x\}, \tag{36}$$

and the corresponding inverse cumulative density function $\widehat{\Phi}_L^{-1}(\cdot)$:

$$\widehat{\Phi}_L^{-1}(\kappa) = \min\left\{z : \frac{1}{L}\sum_{i=1}^{L}\mathbb{I}\{d_i \le z\} \ge \kappa\right\}, \tag{37}$$

where $\kappa \in [0,1]$. The following Dvoretzky–Kiefer–Wolfowitz lemma shows the concentration between $\widehat{\Phi}_L(a)$ and $\Phi(a)$ for any $a \in \mathbb{R}$.

Lemma C.1. *(Dvoretzky–Kiefer–Wolfowitz lemma) Let $\{d_i\}_{i=1}^T$ be T i.i.d. samples drawn from distribution $\mathcal{D}$ with cumulative density function Φ. Define the empirical cumulative density function $\widehat{\Phi}_L(\cdot)$ as shown in Equation (36), $L \in [T]$. Then with probability at least $1-\delta$, for any $x \in \mathbb{R}$,*

$$\left|\widehat{\Phi}_L(x) - \Phi(x)\right| \le \sqrt{\frac{1}{2L}\log\frac{2}{\delta}}.$$

Moreover, by applying a union bounded over all $L \in [T]$, with probability at least $1-\delta$, for any $x \in \mathbb{R}$ and $L' \in [T]$, it holds that

$$\left|\widehat{\Phi}_{L'}(x) - \Phi(x)\right| \le \sqrt{\frac{1}{2L'}\log\frac{2T}{\delta}}.$$

The next lemma shows the stability of $\widehat{\Phi}_L^{-1}(\cdot)$ on consecutive grids of length $\frac{1}{T}$ over $[0,1]$, which turns out to be important to prove our main lemma Lemma C.3.

Lemma C.2. *Let $\{d_i\}_{i=1}^T$ be T i.i.d. samples from distribution $\mathcal{D}$ satisfying Assumption 2. Let $\widehat{\Phi}_L(\cdot)$ be the empirical cumulative density function constructed by $\{d_i\}_{i=1}^L$ as shown in Equation (36). The inverse of the empirical density function $\widehat{\Phi}^{-1}(\cdot)$ is defined in Equation (37). Then with probability at least $1-\delta$, for any $\kappa \in \left\{\frac{i}{T}\right\}_{i=0}^{T-1}$ and any $L \in [T]$,*

$$\widehat{\Phi}_L^{-1}\left(\kappa + \frac{1}{T}\right) - \widehat{\Phi}_L^{-1}(\kappa) \le \frac{2}{\gamma L}\log\frac{LT}{\delta}.$$

Proof. Fix any $L \in [T]$. Without loss of generality, we assume that $d_1 \leq d_2 \leq \ldots \leq d_L$ be the ordered realized demand and let $d_0 = 0$. According to the definition of $\widehat{\Phi}_L^{-1}$, we know that for $a \in (\frac{i-1}{L}, \frac{i}{L}]$, $i \in [L]$

$$\widehat{\Phi}_L^{-1}(a) = d_i.$$

Moreover, as $L \leq T$, meaning that $[\kappa, \kappa + \frac{1}{T}] \subseteq (\frac{i-1}{L}, \frac{i+1}{L}]$ for some $i \in [L]$, we know that for any $\kappa \in \left\{\frac{i}{T}\right\}_{i=0}^{T-1}$,

$$\widehat{\Phi}_L^{-1}\left(\kappa + \frac{1}{T}\right) - \widehat{\Phi}_L^{-1}(\kappa) \leq \max_{i \in [L]} |d_i - d_{i-1}|.$$

Note that according to the property of cumulative density function, $\Phi(x)$ with $x \sim \mathcal{D}$ follows the uniform distribution $\mathcal{U}[0,1]$. According to the property of the ordered statistics of $\mathcal{U}[0,1]$, let $\Delta_k = \Phi(d_k) - \Phi(d_{k-1})$ be the gap between the $k-1$-th and the k-th ordered statistics, $k \in [L]$ and we have Δ_k follows the Beta distribution $\Delta_k \sim \text{Beta}(1, L)$. Therefore, for any $k \in [L]$,

$$\mathcal{P}\left[\Delta_k \geq r\right] = \int_r^1 L(1-u)^{L-1} du = (1-r)^L.$$

Let $r = \frac{1}{L} \log \frac{L}{\delta}$. Then we have

$$\begin{aligned}
\mathcal{P}\left[\exists k \in [L], \Delta_k \geq r\right] &\leq \sum_{k=1}^{L} \mathcal{P}\left[\Delta_k \geq r\right] \\
&= L(1-r)^L \\
&\leq L\left(1 - \frac{1}{L}\log\frac{L}{\delta}\right)^L \\
&\leq L\left(\left(1 - \frac{1}{L}\log\frac{L}{\delta}\right)^{\frac{L}{\log\frac{L}{\delta}}}\right)^{\log\frac{L}{\delta}} \\
&\leq \delta.
\end{aligned}$$

Therefore, with probability at least $1-\delta$, we have $\Delta_k \leq \frac{1}{L}\log\frac{L}{\delta}$, for all $k \in [L]$. According to the assumption that $\phi(d) \geq \gamma$, we have with probability $1-\delta$,

$$\max_{i \in [L]} |d_i - d_{i-1}| \leq \max_{i \in [L]} \frac{1}{\gamma} \cdot |\Phi(d_i) - \Phi(d_{i-1})| \leq \frac{1}{\gamma L} \log \frac{L}{\delta}. \tag{38}$$

Taking a union bound over all possible choices of κ and $L \in [T]$ gives the conclusion. □

Now we are ready to prove Lemma C.3. Note that different from the concentration result which holds for a specific known κ (e.g. Proposition 3 in (Chen et al., 2021)), with the help of Lemma C.2, Lemma C.3 proves that with high probability, for *all* $\kappa \in [0,1]$, we have the difference between $\widehat{\Phi}_L^{-1}(\kappa)$ and $\Phi^{-1}(\kappa)$ bounded by $\widetilde{\mathcal{O}}(1/\sqrt{L})$, which is what we require in the decentralized setting with contract in the coupling model introduced in Section 3 as $\frac{\beta_m}{\beta_m + h_2}$ can take arbitrary values between $[0,1]$.

Lemma C.3. *Let $\{d_i\}_{i=1}^T$ be T i.i.d. samples from distribution $\mathcal{D}$ which satisfies Assumption 2. Let $\widehat{\Phi}_L(\cdot)$ be the empirical cumulative density function constructed by $\{d_i\}_{i=1}^L$ as shown in Equation (36). The inverse of the empirical density function $\widehat{\Phi}^{-1}(\cdot)$ is defined in Equation (37). Then with probability at least $1-\delta$, for any $\kappa \in [0,1]$ and any $L \in [T]$, it holds that*

$$\left|\widehat{\Phi}_L^{-1}(\kappa) - \Phi^{-1}(\kappa)\right| \leq C_0 \sqrt{\frac{\log\frac{TD}{\delta}}{L}},$$

$$\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa\right| \leq C_0 \sqrt{\frac{\log\frac{TD}{\delta}}{L}},$$

where $C_0 > 0$ are some universal constants.

Proof. For any fixed $\kappa \in \{\frac{i}{T}\}_{i=0}^{T}$ and $L \in [T]$, we know that

$$\begin{aligned}
&\mathcal{P}\left[\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa \le -\xi\right] \\
&\le \mathcal{P}\left[\widehat{\Phi}_L^{-1}(\kappa) \le \Phi^{-1}(\kappa - \xi)\right] \\
&\le \mathcal{P}\left[\frac{1}{L}\sum_{i=1}^{L}\left\{d_i \le \Phi^{-1}(\kappa-\xi)\right\} \ge \kappa\right] && \text{(according to the definition of } \widehat{\Phi}_L^{-1}(\cdot)) \\
&\le \mathcal{P}\left[\frac{1}{L}\sum_{i=1}^{L}\left\{d_i \le \Phi^{-1}(\kappa-\xi)\right\} - (\kappa-\xi) \ge \xi\right] \le \exp(-2L\xi^2), && (39)
\end{aligned}$$

where the last inequality is by Hoeffding's inequality. On the other hand,

$$\begin{aligned}
&\mathcal{P}\left[\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa \ge \xi\right] \\
&\le \mathcal{P}\left[\widehat{\Phi}_L^{-1}(\kappa) \ge \Phi^{-1}(\kappa + \xi)\right] \\
&\le \mathcal{P}\left[\frac{1}{L}\sum_{i=1}^{L}\left\{d_i \le \Phi^{-1}(\kappa+\xi)\right\} < \kappa\right] && \text{(according to the definition of } \widehat{\Phi}_L^{-1}(\cdot)) \\
&\le \mathcal{P}\left[\frac{1}{L}\sum_{i=1}^{L}\left\{d_i \le \Phi^{-1}(\kappa+\xi)\right\} - (\kappa+\xi) < -\xi\right] \le \exp(-2L\xi^2),
\end{aligned}$$

Therefore, we conclude that

$$\mathcal{P}\left[\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa\right| \ge \xi\right] \le 2\exp(-2L\xi^2).$$

Therefore, with probability at least $1-\delta$,

$$\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa\right| \le \sqrt{\frac{\log\frac{2}{\delta}}{2L}}.$$

Taking a union bound over all $\kappa \in \{\frac{i}{T}\}_{i=0}^{T}$, with probability at least $1-\delta$, we can obtain that for all $\kappa \in \{\frac{i}{T}\}_{i=0}^{T}$,

$$\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa\right| \le \sqrt{\frac{\log\frac{2TD}{\delta}}{2L}}. \tag{40}$$

Next, for any $\kappa \in [0,1]$, let $\kappa_0 \ge \kappa, \kappa_1 \le \kappa$ be the real number such that $\kappa_0 - \kappa$ and $\kappa - \kappa_1$ is minimized and $\kappa_0, \kappa_1 \in \{\frac{i}{T}\}_{i=0}^{T}$, $\kappa_0 - \kappa_1 = \frac{1}{T}$. Then, according to Lemma C.2, with probability at least $1 - \frac{\delta}{2}$, we have

$$\begin{aligned}
&\left|\widehat{\Phi}_L^{-1}(\kappa) - \Phi^{-1}(\kappa)\right| \\
&= \left|\widehat{\Phi}_L^{-1}(\kappa) - \widehat{\Phi}_L^{-1}(\kappa_1) + \widehat{\Phi}_L^{-1}(\kappa_1) - \Phi^{-1}(\kappa_1) + \Phi^{-1}(\kappa_1) - \Phi^{-1}(\kappa)\right| \\
&\le \left|\widehat{\Phi}_L^{-1}(\kappa) - \widehat{\Phi}_L^{-1}(\kappa_1)\right| + \left|\widehat{\Phi}_L^{-1}(\kappa_1) - \Phi^{-1}(\kappa_1)\right| + \left|\Phi^{-1}(\kappa_1) - \Phi^{-1}(\kappa_0)\right| && (\Phi^{-1}(\kappa_1) \le \Phi^{-1}(\kappa) \le \Phi^{-1}(\kappa_0)) \\
&\le \left|\widehat{\Phi}_L^{-1}(\kappa_0) - \widehat{\Phi}_L^{-1}(\kappa_1)\right| + \left|\widehat{\Phi}_L^{-1}(\kappa_1) - \Phi^{-1}(\kappa_1)\right| + \left|\Phi^{-1}(\kappa_1) - \Phi^{-1}(\kappa_0)\right| \\
&\le \frac{2}{\gamma L}\log\frac{4LT}{\delta} + \frac{1}{\gamma}\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa_1)\right) - \kappa_1\right| + \left|\Phi^{-1}(\kappa_1) - \Phi^{-1}(\kappa_0)\right| && \text{(Lemma C.2)} \\
&\le \frac{1}{\gamma L}\log\frac{LT}{\delta} + \frac{1}{\gamma}\sqrt{\frac{\log\frac{8TD}{\delta}}{2L}} + \frac{1}{\gamma T} && \text{(Equation (40))} \\
&\le C'\sqrt{\frac{\log\frac{TD}{\delta}}{L}},
\end{aligned}$$

where $C' > 0$ is some universal constant.

For $\left|\Phi(\widehat{\Phi}_L^{-1}(\kappa)) - \kappa\right|$, define κ_0 and κ_1 the same as before and we know that with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \kappa\right| \\
&= \left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa)\right) - \Phi\left(\widehat{\Phi}_L^{-1}(\kappa_1)\right) + \Phi\left(\widehat{\Phi}_L^{-1}(\kappa_1)\right) - \kappa_1 + \kappa_1 - \kappa\right| \\
&\le \left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa_0)\right) - \Phi\left(\widehat{\Phi}_L^{-1}(\kappa_1)\right)\right| + \left|\Phi\left(\widehat{\Phi}_L^{-1}(\kappa_1)\right) - \kappa_1\right| + |\kappa_1 - \kappa| \\
&\le \Gamma\left|\widehat{\Phi}_L^{-1}(\kappa_0) - \widehat{\Phi}_L^{-1}(\kappa_1)\right| + \sqrt{\frac{\log\frac{2TD}{\delta}}{2L}} + \frac{1}{T} && \text{(Equation (40) and Lipschitzness of } \Phi\text{)} \\
&\le \frac{2\Gamma}{\gamma L}\log\frac{4LT}{\delta} + \sqrt{\frac{\log\frac{8TD}{\delta}}{2L}} + \frac{1}{T} && \text{(Lemma C.2)} \\
&\le C''\sqrt{\frac{\log\frac{TD}{\delta}}{L}},
\end{aligned}$$

where $C'' > 0$ is some universal constant. Taking a union bound over all κ and $L \in [T]$ and choosing $C_0 = \max\{4C', 4C''\}$ finish the proof. □

The last lemma shows the strong convexity of the expectation of the loss function introduced in Equation (8).

Lemma C.4. *For any $h > 0, p > 0$, let $f(s) = h\mathbb{E}_{x\sim\mathcal{D}}\left[(s-x)^+\right] + p\mathbb{E}_{x\sim\mathcal{D}}\left[(s-x)^-\right]$ where $\mathcal{D}$ satisfies Assumption 1 and Assumption 2. Then $f(s)$ is strongly convex in s with strongly convex parameter $\sigma = (h+p)\gamma$ where γ is defined in Assumption 2.*

Proof. Taking the second order gradient of $f(s)$, we know that

$$\nabla^2 f(s) = (h+p)\phi(x) \ge (h+p)\gamma,$$

where the last inequality is due to Assumption 2. This finishes the proof. □

D Experiments

In this section, we show more empirical results for our designed algorithms. Specifically, we verify the empirical performance of Algorithm 1 and Algorithm 3 in our model. We construct different bounded demand distributions listed as follows: 1) Gaussian distribution $\mathcal{N}(3,1)$ clipped on support $[1,4]$; 2) uniform distribution over $[1,4]$; 3) exponential distribution with mean 3 clipped on support $[1,4]$. We also set the number of round to be $T = 800000$ and choose the cost configuration to be $(h_1, h_2, p_1) = \{(0.3, 0.1, 0.5), (0.4, 0.25, 0.6), (0.5, 0.35, 0.75), (0.6, 0.4, 0.85)\}$. For each demand distribution, 128 trials are processed and we calculate the mean and the standard deviation of the regret over the 128 trials. The results are shown in Figure 1. The results show the effectiveness of our proposed algorithm in both the centralized and decentralized setting.

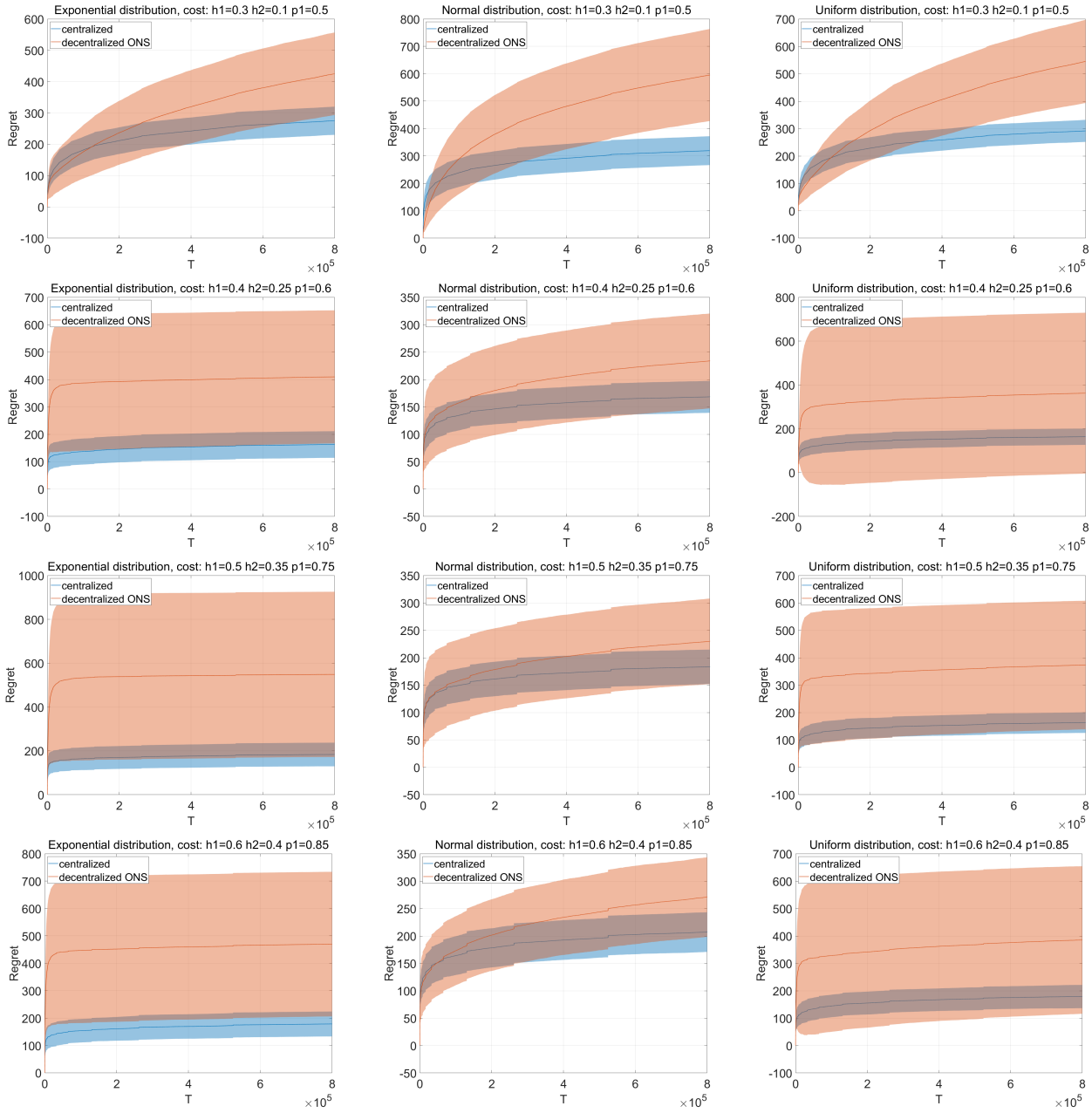

Figure 1: Empirical results of our algorithms applied to our model with cost parameters $(h_1, h_2, p_1) = \{(0.3, 0.1, 0.5), (0.4, 0.25, 0.6), (0.5, 0.35, 0.75), (0.6, 0.4, 0.85)\}$ and $T = 800000$. Each column shows the results of a specific demand distribution with different cost parameter configurations. The algorithm is processed over 128 trials of demand sequences drawn from the four distributions. The solid line is the mean over 128 trials and the shaded area is mean $\pm$ std. The performance of of Algorithm 1 is shown in the blue curve ("centralized") and the one of Algorithm 3 is shown in the orange curve ("decentralized ONS"). The results show the effectiveness of our proposed algorithms.