Joint Pricing and Resource Allocation: An Optimal Online-Learning Approach

Jianyu Xu
Carnegie Mellon University
jianyux@andrew.cmu.edu

Xuan Wang
HKUST
xuanwang@ust.hk

Yu-Xiang Wang
UC San Diego
yuxiangw@ucsd.edu

Jiashuo Jiang
HKUST
jsjiang@ust.hk

Abstract

We study an online learning problem on dynamic pricing and resource allocation, where we make joint pricing and inventory decisions to maximize the overall net profit. We consider the stochastic dependence of demands on the price, which complicates the resource allocation process and introduces significant non-convexity and non-smoothness to the problem. To solve this problem, we develop an efficient algorithm that utilizes a "Lower-Confidence Bound (LCB)" meta-strategy over multiple OCO agents. Our algorithm achieves $\tilde{O}(\sqrt{Tmn})$ regret (for m suppliers and n consumers), which is *optimal* with respect to the time horizon T. Our results illustrate an effective integration of statistical learning methodologies with complex operations research problems.

1 Introduction

The problem of *dynamic pricing* examines strategies of setting and adjusting prices in response to varying customer behaviors and market conditions. The mainstream of existing works on dynamic pricing, including Kleinberg and Leighton (2003); Broder and Rusmevichientong (2012); Cohen et al. (2020), focuses on the estimation of *demand* curves while putting aside the decisions on the *supply* side. Another series of literature, including Besbes and Zeevi (2009); Chen et al. (2019); Keskin et al. (2022), takes supply and inventories into account. However, these works simplify the supply cost as uniform and static, underestimating the difficulty of allocating products through sophisticated supply chains among multiple parties such as factories, warehouses, and retailers.

On the other hand, the problem of *resource allocation* – to serve different demand classes with various types of resources – presents a complex challenge within the field of operations research. Analogous to online dynamic pricing, the recent proliferation of e-platforms has magnified the importance of developing online allocation algorithms that efficiently manage supply and demand on the fly while maximizing cumulative utilities. However, traditional approaches in resource allocation are insufficient in depicting scenarios where the demand is stochastic and dependent on the price (Hwang et al., 2021). Therefore, it is critical to develop price-dependent online allocation models and methodologies that can simultaneously learn the demand curve and optimize the joint decisions on price, inventory, and allocation.

Problem Setting. This work introduces a novel framework for tackling the online pricing and allocation problem with an emphasis on learning under uncertainty. More specifically, we consider a problem setting where both the price and inventory decisions are made at the beginning of each time

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

period, followed by inventory allocation based on the realized price-dependent stochastic demands during that period. We summarize the proposed framework as follows:

Pricing and Allocation. For $t = 1, 2, ..., T$:

1. Determine the inventories of m suppliers as $\vec{I} := [I_1, I_2, \ldots, I_m]^\top$ and incur an immediate inventory cost $\sum_{i=1}^m \gamma_i \cdot I_i$.
2. Propose a price p_t for all n consumers.
3. Based on the price p_t, consumers generate their demands as $\vec{D} := [D_1, D_2, \ldots, D_n]^\top$.
4. We allocate inventories $\vec{I}$ to satisfy demands $\vec{D}$. The allocation from Supplier i to Consumer j is denoted as $X_{i,j}$. The total supplying cost is $\sum_{i=1}^m \sum_{j=1}^n C_{i,j} \cdot X_{i,j}$.
5. We receive a payment from the consumers as $\sum_{i=1}^m \sum_{j=1}^n p_t \cdot X_{i,j}$ in total.

We assume that the inventories are *perishable* and the leftover inventory cannot be carried over to the following period. Furthermore, we assume that the price is *identical* for all consumers. We formalize the above process as solving the following two-stage stochastic programming problem:

$$\min_{p,\vec{I}} \langle \vec{\gamma}, \vec{I} \rangle + \mathbb{E}_{\vec{D}}[g(\vec{I}, p, \vec{D})|p].$$

Here $g(\vec{I}, p, \vec{D})$ is the minimum *negative net profit* (or *loss*) under the best allocation scheme given inventories $\vec{I}$, price p, and demand $\vec{D}$. A rigorous problem setup is presented in Appendix B.

$$\begin{aligned} g(\vec{I}, p, \vec{D}) = \min_{\mathbf{X} \in \mathbb{R}_+^{m \times n}} & \sum_{i=1}^m \sum_{j=1}^n (-p + C_{i,j}) X_{i,j} \\ \text{s.t.} \quad & X_{i,j} \geq 0, \forall i \in [m], j \in [n] \\ & \sum_{i=1}^m X_{i,j} \leq D_j, \forall j \in [n]; \quad \sum_{j=1}^n X_{i,j} \leq I_i, \forall i \in [m]. \end{aligned} \tag{1}$$

Assumptions. We make a few technical assumptions to clarify the scope of our methods.

1. The realized demand is *linear* and *noisy*: $\vec{D}_t = \vec{a} - \vec{b}p + \vec{N}_t$ where $\vec{a}, \vec{b}, \vec{N}_t \in \mathbb{R}^n$ are the base demand, the price sensitivity, and an i.i.d. zero-mean market noise.
2. The norms of parameters $\vec{\gamma}, \vec{a}, \vec{b}, \vec{I}$ and p are *upper-bounded* by *known* constants.
3. We have *full* knowledge on the problem parameters $\vec{\gamma}$ and $\{C_{i,j}\}$. We *do not* know the model parameters $\vec{a}, \vec{b}$ or the distribution of $\vec{N}_t$.

Goals and Metrics. Given the linear demand model, we define the *cost* function to minimize

$$Q_t(\vec{I}, p) := \langle \vec{\gamma}, \vec{I} \rangle + g(\vec{I}, p, \vec{a} - \vec{b} \cdot p + \vec{N}_t), \quad Q(\vec{I}, p) := \mathbb{E}_{\vec{N}_t}[Q_t(\vec{I}, p)]. \tag{2}$$

We adopt **regret** as a performance metric:

$$Reg := \sum_{t=1}^T Q(\vec{I}_t, p_t) - \min_{\vec{I}, p} Q(\vec{I}, p).$$

1.1 Summary of Results.

In this work, we establish a novel framework for solving the online pricing and allocation problem under demand uncertainties. Our main contributions are twofold:

1. **Algorithmic Design against Non-Convexity**. We propose an efficient online-learning algorithm for the (price,inventory) joint decision, which is highly *non-convex*. To navigate to the global optimal decisions among many sub-optimals, we incorporate an optimistic meta-algorithm to manage multiple OCO agents working locally.
2. **Regret Analysis**. We show that our algorithm achieves $\tilde{O}(\sqrt{Tmn})$ regret, which is *optimal* with respect to T as it matches the information-theoretic lower bound.

To the best of our knowledge, we are the first to study dynamic pricing and inventory control under the framework of online resource allocation with uncertainty. Wang et al. (2021b); Chen and Gallego (2021) also study related topics, but their approaches cannot overcome local convexity, multiple suboptimals, or non-smoothness in our setting, which we develop new techniques to solve.

2 Algorithm

Here we propose an online learning algorithm that makes evolving (inventory, price) decisions.

Algorithm 1 LCB Meta Algorithm

Input: m, n, T, supply costs $\{C_{i,j}\}_{i=1,j=1}^{m,n}$, parameters δ_K, n_0.
Sorting: $\{C_{i,j}\}_{i=1,j=1}^{m,n} \Rightarrow \{C_{i_k,j_k}\}_{k=0}^{mn+1}$ according to Algorithm 2.
Initialization: $\hat{W}_K := 0, \Delta_K := +\infty, LCB_K := -\infty, K = 0, 1, 2, \ldots, mn$.
for $K = 0, 1, 2, \ldots, mn$ **do**
  Initialize local agent $\mathcal{A}_K$ as follows:
  Let $L_{K,1} = C_{i_K, j_K}, U_{K,1} = C_{i_{K+1}, j_{K+1}}$,
  and $a_{K,1} = \frac{3L_{K,1}+U_{K,1}}{4}, c_{K,1} = \frac{L_{K,1}+U_{K,1}}{2}, b_{K,1} = \frac{L_{K,1}+3U_{K,1}}{4}$.
  for $p = a_{K,1}, b_{K,1}, c_{K,1}$, **do**
    Propose inventory $\vec{I_0} := [1, 1, \ldots, 1]^\top$ and price p.
    Record the marginal function of $\vec{I}$ as $Q_0(\vec{I}, p)$.
    Find $I_{K,1,0}(p) := \text{argmin}_{\vec{I}} Q_0(\vec{I}, p)$.
  end for
  Set the Stage Flag $\pounds_K \leftarrow 1$ for Agent $\mathcal{A}_K$.
end for
while $t \leq T$ **do**
  Let $\hat{K} := \text{argmin}_{K \in [mn] \cup 0} LCB_K$.
  if $\pounds_{\hat{K}} == 1$ **then**
    Run Algorithm 3 (Stage 1 of $\mathcal{A}_{\hat{K}}$) for *one sub-epoch*.
    $t += |$the length of this sub-epoch$|$.
    Update $\hat{W}_{\hat{K}}, \Delta_{\hat{K}}$ and $\pounds_{\hat{K}}$ according to the statement of $\mathcal{A}_{\hat{K}}$.
  else if $\pounds_{\hat{K}} == 2$ **then**
    Run Algorithm 4 (Stage 2 of $\mathcal{A}_{\hat{K}}$) until its *completion*.
    $t += |$the length of this Stage 2$|$.
    Update $\hat{W}_{\hat{K}}, \Delta_{\hat{K}}$ according to the statement of $\mathcal{A}_{\hat{K}}$.
    Update $\pounds_{\hat{K}} \leftarrow 3$.
  else if $\pounds_{\hat{K}} == 3$ **then**
    Run Algorithm 5 (Stage 3 of $\mathcal{A}_{\hat{K}}$) for *one single time period*.
    $t += 1$.
    Update $\hat{W}_{\hat{K}}, \Delta_{\hat{K}}$ according to the statement of $\mathcal{A}_{\hat{K}}$.
  end if
  for $K' = 0, 1, 2, \ldots, mn$. **do**
    Update $LCB_{K'} \leftarrow \hat{W}_{K'} - 34 \cdot \Delta_{K'}$.
  end for
end while

2.1 Overview: Local (Vertical) and Interval (Horizontal) Uncertainty Contractions

Our setting poses two main challenges: *bandit feedback* and *non-convexity* with respect to price p. Existing dynamic pricing methods address these with continuum-armed bandit algorithms, but these typically yield $O(T^{2/3})$ regret (Kleinberg, 2004), which is not optimal in this scenario. To improve upon this, we leverage the *piecewise-convex* structure of the expected loss: The domain $[0, p_{\max}]$ can be divided into intervals, in each of which the function is locally convex. We assign a dedicated *agent* to each interval, applying online convex optimization (OCO) to approach its local optimal. We introduce twofold confidence bounds for each agent:

(a) **Vertical confidence bound**: This quantifies the uncertainty in the agent's estimate of the local optimal value within its convex interval. It is updated at the end of each *sub-epoch*. (We provide the definitions of sub-epoch and epoch in Appendix C).

(b) **Horizontal confidence bound**: This describes the shrinking search space within the interval where the local optimum is likely to be found. It is updated at the end of each *epoch* (and is updated at every t if the agent reaches Stage 3).

Algorithm 2 Sort tuples $\{(i,j)\}_{i=1,j=1}^{i=m,j=n}$ as follows:

Input: $\{(i,j)\}$ tuples.
for each different pairs of tuples (i,j) and (i',j') **do**
  If $C_{i,j} < C_{i',j'}$, then $(i,j) \prec (i',j')$.
  If $C_{i,j} = C_{i'.j'}$ and $i > i'$, then $(i,j) \prec (i',j')$.
  If $C_{i,j} = C_{i',j'}$, $i = i'$ and $j > j'$, then $(i,j) \prec (i',j')$.
end for
Output: $\{(i_k, j_k)\}_{k=1}^{mn}$.

As the algorithm proceeds, either the vertical confidence bound (value uncertainty) or the horizontal confidence bound (location uncertainty) for each agent becomes tighter, thereby improving both the accuracy of local optimization and the efficiency of interval selection. This hierarchical design enables us to achieve an improved regret bound of $O(\sqrt{T})$.

Algorithm 1 is the pseudo-code of our algorithm. We firstly sort $\{C_{i,j}\}$ into $\{C_{i_k,j_k}\}_{k=0}^{mn}$ according to Algorithm 2. Then we run each of the $(mn+1)$ agents $\mathcal{A}_K$ for $K = 0, 1, \ldots, mn$ for 3 times. After this warm-up, we select the arg-max agent $\mathcal{A}_{\hat{K}}$ whose Lower Confidence Bound $LCB_{\hat{K}}$ is the minimum, and run $\mathcal{A}_{\hat{K}}$ for a period of time (depending on its stage). The status of each $\mathcal{A}_K$ can be divided into three stages:

(i) In Stage 1 (as Algorithm 3), $\mathcal{A}_K$ searches for the local optimal decision in its domain.
(ii) In Stage 2 (as Algorithm 4), $\mathcal{A}_K$ gathers a sufficient number of samples to ensure an appropriate convergence of its confidence bound.
(iii) In Stage 3 (as Algorithm 5), $\mathcal{A}_K$ purely exploits the local optimal while also updating LCB_K.

In Appendix C, we will introduce each component of our algorithm design (including the meta-algorithm and each OCO agent) in details.

3 Regret Analysis

Here we present our main theorem on the regret bound, along with a demonstration of proof roadmap. We will deliver a complete and rigorous regret analysis in Appendix D.

Theorem 3.1 (Regret). *Algorithm 1 guarantees an $\tilde{O}(\sqrt{Tmn} + mn)$ regret with probability at least $1-\epsilon$. Here $\tilde{O}(\cdot)$ omits the dependence on $\log\frac{1}{\epsilon}$ and $\log T$.*

This regret rate is near-optimal with respect to T, as it matches the information-theoretic lower bound of $\Omega(\sqrt{T})$ proved by Broder and Rusmevichientong (2012). Here we present a proof sketch:

Proof Sketch. Define a concept *Sup-Regret* for each local agent $\mathcal{A}_K$, measuring the performance regret of $\mathcal{A}_K$ with respect to its local optimal decision $(\vec{I}_K^*, p_K^*)$. We have the following propositions

1. **Horizontal convergence:** For agent $\mathcal{A}_K$ that has been running for T_K time periods so far, its sub-regret is bounded by $\tilde{O}(\sqrt{T_K})$.

2. **Vertical convergence:** For agent $\mathcal{A}_K$ running for T_K time periods, its *vertical* confidence bound satisfies $\Delta_K = \tilde{O}(\frac{1}{\sqrt{T_K}})$ for sufficiently large T_K. Therefore, after a sufficiently long time $N_0 = \tilde{O}(mn)$, we have $\Delta_K < +\infty, \forall K = 0, 1, \ldots, mn$.

Given the propositions listed above, we may upper-bound the total regret by

$$\begin{aligned}\text{Reg} &= \sum_{K=0}^{mn} \text{Sub-Reg}_K + \sum_{t=1}^{T} Q(\vec{I}_{K_t}^*, p_{K_t}^*) - Q(\vec{I}^*, p^*) \\ &= \tilde{O}(\sum_{K=0}^{mn} \sqrt{T_K} + \sum_{t=N_0}^{T} \Delta_{K_t} + N_0) = \tilde{O}(\sqrt{mn \cdot T} + mn).\end{aligned} \tag{3}$$

■

References

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. (2011). Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24.

Asadpour, A., Wang, X., and Zhang, J. (2020). Online resource allocation with limited flexibility. *Management Science*, 66(2):642–666.

Besbes, O. and Zeevi, A. (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420.

Broder, J. and Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980.

Bumpensanti, P. and Wang, H. (2020). A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science*, 66(7):2993–3009.

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

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

Chen, B., Chao, X., and Wang, Y. (2020). Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research*, 68(5):1445–1456.

Chen, B., Wang, Y., and Zhou, Y. (2023). Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. *Management Science*.

Chen, N. and Gallego, G. (2021). Nonparametric pricing analytics with customer covariates. *Operations Research*.

Chen, Q., Jasin, S., and Duenyas, I. (2021b). Joint learning and optimization of multi-product pricing with finite resource capacity and unknown demand parameters. *Operations Research*, 69(2):560–573.

Choi, Y.-G., Kim, G.-S., Choi, Y., Cho, W., Paik, M. C., and Oh, M.-h. (2023). Semi-parametric contextual pricing algorithm using cox proportional hazards model. In *International Conference on Machine Learning*, pages 5771–5786. PMLR.

Cohen, M. C., Lobel, I., and Paes Leme, R. (2020). Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943.

Cournot, A. A. (1897). *Researches into the Mathematical Principles of the Theory of Wealth*. Macmillan.

Ferreira, K. J., Simchi-Levi, D., and Wang, H. (2018). Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602.

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

Hwang, D., Jaillet, P., and Manshadi, V. (2021). Online resource allocation under partially predictable demand. *Operations Research*, 69(3):895–915.

Jasin, S. and Kumar, S. (2012). A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research*, 37(2):313–345.

Jenatton, R., Huang, J., and Archambeau, C. (2016). Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pages 402–411. PMLR.

Jia, H., Shi, C., and Shen, S. (2022). Online learning and pricing with reusable resources: Linear bandits with sub-exponential rewards. In *International Conference on Machine Learning*, pages 10135–10160. PMLR.

Jiang, J., Ma, W., and Zhang, J. (2022). Degeneracy is ok: Logarithmic regret for network revenue management with indiscrete distributions. *arXiv preprint arXiv:2210.07996*.

Keskin, N. B., Li, Y., and Song, J.-S. (2022). Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Science*, 68(3):1938–1958.

Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17:697–704.

Kleinberg, R. and Leighton, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symposium on Foundations of Computer Science (FOCS-03)*, pages 594–605. IEEE.

LaFrance, J. T. (1985). Linear demand functions in theory and practice. *Journal of Economic theory*, 37(1):147–166.

Leme, R. P., Sivan, B., Teng, Y., and Worah, P. (2021). Learning to price against a moving target. In *International Conference on Machine Learning*, pages 6223–6232. PMLR.

Reiman, M. I. and Wang, Q. (2008). An asymptotically optimal policy for a quantity-based network revenue management problem. *Mathematics of Operations Research*, 33(2):257–282.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Simchi-Levi, D. and Wang, C. (2023). Pricing experimental design: causal effect, expected revenue and tail risk. In *International Conference on Machine Learning*, pages 31788–31799. PMLR.

Vera, A. and Banerjee, S. (2019). The bayesian prophet: A low-regret framework for online decision making. *ACM SIGMETRICS Performance Evaluation Review*, 47(1):81–82.

Vera, A., Banerjee, S., and Gurvich, I. (2021). Online allocation and pricing: Constant regret via bellman inequalities. *Operations Research*, 69(3):821–840.

Wang, H., Talluri, K., and Li, X. (2021a). On dynamic pricing with covariates. *arXiv preprint arXiv:2112.13254*.

Wang, Y., Chen, B., and Simchi-Levi, D. (2021b). Multimodal dynamic pricing. *Management Science*.

A Related Works

In this section, we present a review of the existing literature on pricing and allocation problems.

Dynamic Pricing. Quantitative research on dynamic pricing dates back to Cournot (1897) and has attracted significant attention in the field of machine learning (Leme et al., 2021; Jia et al., 2022; Choi et al., 2023; Simchi-Levi and Wang, 2023). For single-product pricing problems, the crux is to learn the demand curve and approach the optimal price. Under the assumptions of bandit feedback and k^{th}-smooth demand curves, Wang et al. (2021a) achieves an $O(T^{\frac{k+1}{2k+1}})$ regret. However, their methodologies are not applicable to our setting: Our objective function is only Lipschitz continuous, leading to $k = 1$ and an $O(T^{2/3})$ sub-optimal regret. In contrast, the piecewise convex property in our problem enables advanced methods to achieve a better regret.

Online Convex Optimization (OCO). OCO models a scenario in which decisions are made iteratively, facing a series of convex loss functions, with the objective of minimizing cumulative regret over time (Shalev-Shwartz et al., 2012) or within certain budgets (Jenatton et al., 2016). In our work, we adopt zeroth-order methods in Agarwal et al. (2011) when iterating within each local convexity interval. For a detailed review of the classic and contemporary results on OCO, we kindly refer the readers to Hazan (2016).

Resource Allocation. There is a broad literature on the study of resource allocation and various policies have been derived under various settings (e.g. Reiman and Wang 2008; Jasin and Kumar 2012; Ferreira et al. 2018; Asadpour et al. 2020; Bumpensanti and Wang 2020; Vera and Banerjee 2019; Jiang et al. 2022). Notably, the intersection of pricing and resource allocation has also been studied, for example, in Chen et al. (2021b) and Vera et al. (2021). However, previous works have primarily focused on either the allocation decision or the pricing decision separately. In contrast, in our paper, we consider a two-stage process where we first make the pricing decision which affects the demand, and then make the allocation decision. This feature distinguishes our paper from previous works on (price-based) resource allocation.

Pricing and Inventory Co-Decisions. The incorporation of inventory constraints into dynamic pricing problems began with the work of Besbes and Zeevi (2009) which assumed a fixed initial stock, and decisions of replenishment were later allowed in Chen et al. (2019). More recent studies, including Chen et al. (2020); Keskin et al. (2022), assumed perishable goods and took inventory costs into account. The stream of work by Chen et al. (2021a, 2023) further assumed the inventory-censoring effect on demands. However, none of them consider the heterogeneity of supply, nor the impact of prices on the allocation process. In our work, we not only model the inventory cost of each warehouse individually, but also depict the unit supplying cost from Warehouse i to Consumer j as a unique coefficient $C_{i,j}$.

B Problem Setup

In this section, we rigorously define the problem we are studying. We firstly formulate the offline version of the problem as a two-stage stochastic program in Appendix B.1. We then develop the formulation of the online version in Appendix B.2, where the demand parameters are unknown. Finally, we present assumptions that are crucial to our online algorithm design by the end of this section.

B.1 Offline Problem Setting

We consider the following scenario where a retail company makes their decisions with the goal of maximizing their net profit. Suppose that the company has m warehouses and n retailers, producing and selling identical products. In general, this company faces the following three problems on inventory, pricing, and allocation:

1. What are the appropriate quantities each warehouse should load?
2. What is the optimal price that retailers should set?
3. How to allocate inventories from warehouses to stores under heterogeneous supply costs?

To address these questions, we model the problem as a two-stage stochastic program.

(i) In Stage 1, the company makes *inventory* decisions $\vec{I} = [I_1, I_2, \ldots, I_m]^\top$, where I_i is the inventory level of Warehouse i. Each unit of inventory at Warehouse i incurs an inventory cost γ_i. In addition, the company decides a uniform *price* p for the products.

(ii) In Stage 2, a stochastic demand $\vec{D} = [D_1, D_2, \ldots, D_n]^\top$ is generated based on the price p, where D_j represents the demand at Retailer j. Then the products are allocated from warehouses to stores in order to fulfill the realized demand. Each unit of supply from Warehouse i to Retailer j incurs an allocation cost $C_{i,j}$, and each unit of fulfilled demand increases the total revenue by p.

The company aims to make the best (inventory, price) joint decisions that maximize their net profit, which can be formulated as the following optimization problem (where we equivalently minimize the negative net profit):

$$\min_{p,\vec{I}} \langle \vec{\gamma}, \vec{I} \rangle + \mathbb{E}_{\vec{D}}[g(\vec{I}, p, \vec{D})|p] \tag{4}$$

where

$$\begin{aligned} g(\vec{I}, p, \vec{D}) = \min_{\mathbf{X} \in \mathbb{R}_+^{m \times n}} & \sum_{i=1}^{m} \sum_{j=1}^{n} (-p + C_{i,j}) X_{i,j} \\ \text{s.t.} \quad & X_{i,j} \geq 0, \forall i \in [m], j \in [n] \\ & \sum_{i=1}^{m} X_{i,j} \leq D_j, \forall j \in [n]; \quad \sum_{j=1}^{n} X_{i,j} \leq I_i, \forall i \in [m]. \end{aligned} \tag{5}$$

Here $X_{i,j}$ represents the quantity of inventories allocated from Warehouse i to Retailer j, and $g(\vec{I}, p, \vec{D})$ is the optimal objective value of the second-stage problem, which optimally allocates inventory $\vec{I}$ to demand $\vec{D}$ based on price p and cost parameters $\{C_{i,j}\}$.

It is worth noting that the distribution of demand $\vec{D}$ is dependent on the price p, and the minimization over p and $\vec{I}$ takes this into account. However, the solution to $g(\vec{I}, p, \vec{D})$ is not relevant to this dependence, as it is solved after the realization of $\vec{D}$.

We denote an optimal solution to Eq. (4) as $(p^*, \vec{I}^*)$, and denote an optimal solution to Eq. (5) as $\mathbf{X}^*$. Since Eq. (5) is a linear programming, we can solve it directly with any standard optimization tool. However, in order to solve Eq. (4), we have to know the distribution of $\vec{D}$ and how it is dependent on the price p, both of which are not directly accessible from the seller's side as they do not have the full knowledge of the entire market. In the next subsection, we will discuss how we can "learn" the demand distribution function under mild assumptions.

Now we propose a lemma that states the marginal convexities of $g(\vec{I}, p, \vec{D})$.

Lemma B.1. *The function $g(\vec{I}, p, \vec{D})$ defined in Eq. (5) is marginally convex on $\vec{I}$ and on $\vec{D}$.*

The key to proving Lemma B.1 is to show that $g(\vec{I}, p, \vec{D})$ is the piecewise maximum over a group of linear functions. Please refer to Appendix F.1 for details.

B.2 Online Problem Setting

Due to insufficient knowledge of the actual demand distribution, the company could propose pairs of $(p, \vec{I})$ that are suboptimal, leading to lower net profits compared to the optimal solution. However, the company has observations on the realized demand at each store, which enables them to estimate demand and subsequently improve their decision-making. In what follows, we study the online decision-making problem of setting prices and managing inventory.

Denote p_t, $\vec{I}_t$ and $\vec{D}_t$ as the price, inventory and realized demand in each time period $t = 1, 2, \ldots, T$, respectively. We make the following semi-parametric assumption on the demand model.

Assumption B.2. Assume the realized demand is *linear* and *noisy*. Specifically, assume

$$\vec{D}_t = \vec{a} - \vec{b} \cdot p + \vec{N}_t. \tag{6}$$

Here $\vec{a}, \vec{b}, \vec{N}_t \in \mathbb{R}^n$ are the base demand, the price sensitivity parameter, and the market noise of the retailers' demand, respectively. Assume $\vec{a}, \vec{b}$ are fixed, and $\vec{N}_t$ are samples drawn from *identical and independent distributions* (i.i.d.) over time t, such that $\mathbb{E}[\vec{N}_t] = \vec{0}$.

Given the linear-and-noisy demand model in (6), we define the cost function that we aim to minimize. Denote

$$
\begin{aligned}
Q(\vec{I}, p) &:= \langle \vec{\gamma}, \vec{I} \rangle + \mathbb{E}_{\vec{N}_t}[g(\vec{I}, p, \vec{a} - \vec{b} \cdot p + \vec{N}_t)] \\
Q_t(\vec{I}, p) &:= \langle \vec{\gamma}, \vec{I} \rangle + g(\vec{I}, p, \vec{a} - \vec{b} \cdot p + \vec{N}_t)
\end{aligned} \tag{7}
$$

Since $g(\vec{I}, p, \vec{D})$ is marginally convex on $\vec{I}$ according to Lemma B.1, we know that $Q(\vec{I}, p)$ and $Q_t(\vec{I}, p)$ are also marginally convex on $\vec{I}$. But what about their marginal behaviors on p? We state in the following lemma:

Lemma B.3. *Sort $\{C_{i,j}\}_{i=1,j=1}^{m,n}$ according to Algorithm 2. Denote $C_{i_0,j_0} = 0$ and $C_{i_{mn+1},j_{mn+1}} = p_{\max}$. For any $K \in \{0\} \cup [mn]$, function $Q_t(\vec{I}, p)$ is Lipschitz and marginally convex on p in range $[C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}]$.*

We defer the proof of Lemma B.3 to Appendix F.2. Furthermore, we have the following results.

Lemma B.4. *Define an optimistic cost function $W(p)$:*

$$
W(p) := \min_{\vec{I} \in \mathbb{R}_+^m} Q(\vec{I}, p). \tag{8}
$$

We have $W(p)$ is L_W-Lipschitz where L_W is a constant. Also, for any $K \in \{0\} \cup [mn]$, the function $W(p)$ is convex in the range $[C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}]$.

The proof of Lemma B.4 is relegated to Appendix F.3. Finally, we define *regret* as the relative loss of net profit compared to that achieved by optimal decisions.

Definition B.5 (Regret)**.** At each time $t = 1, 2, \ldots, T$, denote $\vec{I}_t$ and p_t as inventory and price decisions, respectively. Define

$$
Reg := \sum_{t=1}^{T} Q(\vec{I}_t, p_t) - \min_{\vec{I}, p} Q(\vec{I}, p) \tag{9}
$$

as the regret of decision sequence $\{(\vec{I}_t, p_t)\}_{t=1}^{T}$.

Before we conclude this section, we present some crucial assumptions for our algorithm design.

Assumption B.6 (Boundedness)**.** We assume boundedness on the norms of $\vec{\gamma}, \vec{a}, \vec{b}, \vec{I}$ and on price p. Specifically, there exist constants $\gamma_{\max}, a_{\max}, b_{\max}, I_{\max}, p_{\max}$ such that $\|\vec{\gamma}\|_\infty \leq \gamma_{\max}, \|\vec{a}\|_\infty \leq a_{\max}, \|\vec{b}\|_\infty \leq b_{\max}, \|\vec{I}\|_1 \leq I_{\max}$ and $p \in [0, p_{\max}]$. Without loss of generality, we assume $\gamma_{\max}, a_{\max}, b_{\max}, I_{\max}, p_{\max} \geq 1$.

Assumption B.7 (Knowledge over parameters)**.** We have *full* knowledge on the problem parameters $\vec{\gamma}, \{C_{i,j}\}_{i=1,j=1}^{m,n}$ and the boundedness parameters $\gamma_{\max}, a_{\max}, b_{\max}, I_{\max}, p_{\max}$ before $t = 0$. We do *not* know the model parameters $\vec{a}, \vec{b}$ nor the distribution of $\vec{N}_t$.

C Algorithm Details

In the following subsections, we will introduce each component of our algorithm design (including the meta-algorithm and each OCO agent) in details.

C.1 Meta-Algorithm: a Lower-Confidence-Bound (LCB) Strategy

Since we have full-information feedback over any decision w.r.t. $\vec{I}$, we may always propose greedy inventories without causing bias. However, we only have bandit feedback w.r.t. the price p, as we have no direct feedback on the prices we are not proposing. Therefore, we conduct online learning on the optimistic cost function $W(p) = \min_{\vec{I}} Q(\vec{I}, p)$.

Due to the piecewise convexity of $W(p)$, we divide the price range $[0, p_{\max}]$ into $(mn+1)$ intervals $[C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}], K = 0, 1, \ldots, mn$. Within each interval, we initialize an OCO agent $\mathcal{A}_K$ that is responsible for converging to the local optimal. However, we cannot run multiple OCO agents simultaneously. Therefore, we require a meta-algorithm that serves as a manager over these agents and determine which $\mathcal{A}_K$ to run at each time, so as to locate the optimal price with the least regret.

To achieve this, we develop a lower-confidence-bound (LCB) meta-algorithm as shown in Algorithm 1. We firstly ask each $\mathcal{A}_K$ agent to maintain a confidence bound $[\hat{W}_K - 34\Delta_K, \hat{W}_K + 34\Delta_K]$ of its local optimal. Given this, the meta-algorithm then selects the agent K that *minimizes* the lower confidence bound. As we further show that $\Delta_K \approx O(\sqrt{1/T_K})$ where T_K is the total time periods that $\mathcal{A}_K$ has been running so far, we may upper bound the cumulative regret as $O(\sqrt{Tmn})$.

Algorithm 3 Agent $\mathcal{A}_K$ Stage 1

Obtain $\hat{W}_K, \Delta_K$ and $\pounds_K$ from the Meta-Algorithm (Algorithm 1).
for Epoch $\tau = 1, 2, \ldots, O(\log T)$, **do**
  Let $a_{K,\tau} = \frac{3L_{K,1}+U_{K,1}}{4}, c_{K,\tau} = \frac{L_{K,1}+U_{K,1}}{2}, b_{K,\tau} = \frac{L_{K,1}+3U_{K,1}}{4}$.
  for Sub-epoch $s = 1, 2, \ldots$ **do**
    Let sub-epoch length $n_s := 2^s$
    Define a $flag := 0$ for error-bar update.
    for $\hat{p}_\tau = a_{K,\tau}, b_{K,\tau}, c_{K,\tau}$ **do**
      for $t = 1, 2, \ldots, n_s$ **do**
        Propose decisions $(\vec{I}_t, p_t) = (\vec{I}_{K,\tau,s-1}(\hat{p}_\tau), \hat{p}_\tau)$.
        Observe and record the marginal function $Q_t(\vec{I}, \hat{p}_\tau)$ with respect to $\vec{I}$.
      end for
      Define an aggregated function $Q_{K,\tau,s}(\vec{I}, \hat{p}_\tau) := \frac{1}{n_s} \cdot \sum_{t=1}^{n_s} Q_t(\vec{I}, \hat{p}_\tau)$.
      Define the empirical optimal inventory $I_{K,\tau,s}(\hat{p}_\tau) := \operatorname{argmin}_{\vec{I}} Q_{K,\tau,s}(I, \hat{p}_\tau)$.
      Denote $\hat{Q}_{K,\tau,s,\hat{p}_\tau} := Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau), \hat{p}_\tau)$, and $\Delta_{K,\tau,s} := \frac{\delta_K}{2\sqrt{n_s}}$.
    end for
    if $\hat{Q}_{K,\tau,s,a_{K,\tau}} > \hat{Q}_{K,\tau,s,b_{K,\tau}} + 4\Delta_{K,\tau,s}$ **then**
      Update $L_{K,\tau+1} \leftarrow a_{K,\tau}, U_{K,\tau+1} \leftarrow U_{K,\tau}, flag \leftarrow 1$.
    else if $\hat{Q}_{K,\tau,s,a_{K,\tau}} < \hat{Q}_{K,\tau,s,b_{K,\tau}} - 4\Delta_{K,\tau,s}$ **then**
      Update $L_{K,\tau+1} \leftarrow L_{K,\tau}, U_{K,\tau+1} \leftarrow b_{K,\tau}, flag \leftarrow 1$.
    else if $\hat{Q}_{K,\tau,s,c_{K,\tau}} < \hat{Q}_{K,\tau,s,a_{K,\tau}} - 4\Delta_{K,\tau,s}$ **then**
      Update $L_{K,\tau+1} \leftarrow a_{K,\tau}, U_{K,\tau+1} \leftarrow U_{K,\tau}, flag \leftarrow 1$.
    else if $\hat{Q}_{K,\tau,s,c_{K,\tau}} < \hat{Q}_{K,\tau,s,b_{K,\tau}} - 4\Delta_{K,\tau,s}$ **then**
      Update $L_{K,\tau+1} \leftarrow L_{K,\tau}, U_{K,\tau+1} \leftarrow b_{K,\tau}, flag \leftarrow 1$.
    end if
    if $flag == 1$ **then**
      if $U_{K,\tau+1} - L_{K,\tau+1} > \frac{1}{T}$ **then**
        Continue to Epoch $\tau + 1$ **(without updating** $\hat{W}_K$ **or** Δ_K**).**
      else
        Set $\hat{p}_K^* \leftarrow c_{K,\tau}, \hat{\vec{I}}_{K,0}^* \leftarrow \vec{I}_{K,\tau,s}(c_{K,\tau})$.
        Update $\pounds_K \leftarrow 2$ and **Break (without updating** $\hat{W}_K$ **or** Δ_K**).**
      end if
    else if $\Delta_{K,\tau,s-1} < \Delta_K$ **then**
      Update $\Delta_K \leftarrow \Delta_{K,\tau,s-1}, \hat{W}_K \leftarrow \min_{\hat{p}_\tau \in \{a_{K,\tau}, c_{K,\tau}, b_{K,\tau}\}} \hat{Q}_{K,\tau,s-1,\hat{p}_\tau}$.
    end if
  end for
end for

C.2 Agent $\mathcal{A}_K$: a Zeroth-Order Optimizer

As described in Appendix C.1, we divide the price range $[0, p_{\max}]$ into $(mn+1)$ intervals $[C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}], K = 0, 1, 2, \ldots, mn$, within each of which the objective function $W(p)$ is convex. We then assign an agent $\mathcal{A}_K$ to each interval, conducting online convex optimization (OCO) locally. We require the agent $\mathcal{A}_K$ to learn and converge to the local optimal $p_K^* := \operatorname{argmin}_{p \in [C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}]} W(p)$ over time, while also maintaining a valid error bar $[\hat{W}_K - 34\Delta_K, \hat{W}_K + 34\Delta_K]$ that contains $W(p_K^*)$ with high probability. To achieve the optimal regret, we rely on the following properties of $\mathcal{A}_K$:

Algorithm 4 Agent $\mathcal{A}_K$ Stage 2 (a sub-epoch totally)

Obtain $\hat{W}_K, \varDelta_K$ and $\pounds_K$ from the Meta-Algorithm (Algorithm 1).
Initialization: Set $\hat{p}_K^* := c_{K,\tau}, \hat{\vec{I}}_{K,0}^* = \vec{I}_{K,\tau,s}(c_{K,\tau})$ from Stage 1, and $\check{C}_K := \log_{4/3} p_{\max} + 1$.
if $\varDelta_K < +\infty$ **then**
  Let $N_{K,2} := \frac{4\delta_K^2}{\varDelta_K^2}$
else
  Let $N_{K,2} := n_0 = 6(\log_{4/3} T + \check{C}_K)$
end if
Let $r_K := \log_2(N_{K,2})$
for $r = 1, 2, \ldots, r_K$ **do**
  for $t = 1, 2, \ldots, m_r := 2^r$ **do**
    Propose decisions $(\vec{I}_t, p_t) = (\hat{\vec{I}}_{K,r-1}^*, \hat{p}_K^*)$.
    Observe and record the marginal function $Q_t(\vec{I}, \hat{p}_K^*)$ with respect to $\vec{I}$.
  end for
  Set $\hat{\vec{I}}_{K,r}^* := \text{argmin}_{\vec{I}} \bar{Q}_{K,r}(\vec{I}, \hat{p}_K^*)$, where $\bar{Q}_{K,r}(\vec{I}, \hat{p}_K^*) := \frac{1}{m_r} \cdot \sum_{t=1}^{\top} Q_t(\vec{I}, \hat{p}_K^*)$.
end for
Denote $\vec{I}_K^* := \hat{\vec{I}}_{K,r_K}^*, \hat{Q}_K^* := \bar{Q}_{K,r_K}(\vec{I}_K^*, \hat{p}_K^*)$.
Update $\hat{W}_K \leftarrow \hat{Q}_K^* - L_W \cdot \frac{1}{T}, \varDelta_K \leftarrow \frac{\delta_K}{\sqrt{N_{K,2}}}$ **and** $\pounds_K \leftarrow 3$**.**

Algorithm 5 Agent $\mathcal{A}_K$ Stage 3 (each t as a sub-epoch)

Obtain $\hat{W}_K, \varDelta_K$ and $\pounds_K$ from the Meta-Algorithm (Algorithm 1).
Initialization: Set $\hat{p}_K^*, \hat{\vec{I}}_{K,r_K}^*, \bar{Q}_{K,r_K}(\vec{I}, \hat{p}_K^*)$ from Stage 2 (Algorithm 4).
Denote $\vec{I}_K^* := \hat{\vec{I}}_{K,r_K}^*, \hat{Q}_K^* := \bar{Q}_{K,r_K}(\vec{I}_K^*, \hat{p}_K^*)$.
Let $N_{K,3} \leftarrow N_{K,2}$ as its initialization.
while $t \leq T$ **do**
  Propose decisions $(\vec{I}_t, p_t) = (\vec{I}_K^*, \hat{p}_K^*)$.
  Observe and record the marginal function $Q_t(\vec{I}, \hat{p}_K^*)$ with respect to $\vec{I}$.
  Update $\hat{Q}_K^* \leftarrow \frac{N_{K,3} \cdot \hat{Q}_K^* + Q_t(\vec{I}_t, \hat{p}_K^*)}{N_{K,3}+1}$, and $N_{K,3} \leftarrow N_{K,3} + 1$
  Update $\hat{W}_K \leftarrow \hat{Q}_K^* - L_W \cdot \frac{1}{T}$ **and** $\varDelta_K \leftarrow \frac{\delta_K}{\sqrt{N_{K,3}}}$**.**
end while

(a) The cumulative sub-regret of $\mathcal{A}_K$, i.e. performance suboptimality compared with $W(p_K^*)$, is bounded by $\tilde{O}(\sqrt{T_K})$ as an optimal rate of OCO (if we have run $\mathcal{A}_K$ for T_K times so far).

(b) The error bar $\varDelta_K$ is bounded by $\tilde{O}(\sqrt{\frac{1}{T_K}})$ as a requirement of the meta-algorithm.

Here we elaborate each component of $\mathcal{A}_K$'s algorithmic design in detail.

Horizontal search space for p_K^*. In the design of Stage 1 algorithm as presented in Algorithm 3, we adopt the framework of zeroth-order online convex optimization. Specifically, we establish an epoch-based update rule of the search space of local optimal p_K^*. The search space (interval) for Epoch $\tau = 1, 2, \ldots$ is denoted as $[L_{K,\tau}, U_{K,\tau}]$. Within each epoch, we divide the time horizon into a series of *doubling sub-epochs* to gather samples for $W(a), W(b), W(c)$ where a, b, c are the three quarter points. By the end of each sub-epoch, we update the estimates and examine whether their estimation error bar is *separable* according to certain rules. As we keep doubling the size of sub-epochs, the estimation error bars are shrinking exponentially until they are separated. Then we reduce the search space by one quarter and proceed to Epoch $\tau + 1$ Sub-Epoch 1. When the search space is as sufficiently small as $O(1/T)$, we stop searching and proceed to Stage 2.

Vertical uncertainty bound for $W(p_K^*)$. In Stage 1, we maintain an error bar Δ_K as the confidence bound of estimating each local optimal $W(p_K^*)$. We show that the error bar has a size of $\tilde{O}(\frac{1}{\sqrt{T_K}})$ if we have run $\mathcal{A}_K$ for T_K times so far. In addition to the statistical concentrations, another intuition of this fact comes from Lemma F.5: A not-distinguishable situation implies a comparable uncertainty bound for the optimal.

Complementary sampling to enhance Δ_K. It is worth noting that $\Delta_K < +\infty$ does not exist for granted even after Stage 1. This is because we cannot update Δ_K when the search space $[L_K, U_K]$ is updated, i.e., no simultaneous "horizontal converging" and "vertical converging". As a consequence, if we are very "lucky" that we can always reduce the search space in the *first* sub-epoch of every epoch until $U_{K,\tau} - L_{K,\tau} \leq 1/T$, then we will have $\Delta_K = +\infty$ until Stage 2. We resolve this issue in two approaches: (1) We upper bound the time periods before any $\Delta_K < +\infty$ by $O(\log T)$ based on the Pigeon-Hole Theorem. (2) We have Stage 2 as a complementary sampling stage without causing excessive regret. By the end of Stage 2, we will have an ideal error bar for each agent $\mathcal{A}_K$.

Pure local exploitation contributing to global LCB. From Agent $\mathcal{A}_K$'s perspective, it runs pure exploitation in Stage 3 (Algorithm 5) without causing extra sub-regret. However, it still keeps updating the estimates of $\hat{W}_K$ and Δ_K to facilitate the LCB meta-algorithm.

Technical Novelty We propose a unique methodology undergoing "horizontal-and-vertical" convergence simultaneously, for the first time. In contrast, existing works adopt either "vertical convergence" such as bandits algorithms (which allow non-convexity of objective functions but cannot achieve $O(\sqrt{T})$ regret even with smoothness assumptions), or "horizontal convergence" which is applicable to many online planning and optimization scenarios but requires global convexity assumptions.

D Regret Analysis Details

In this section, we provide the theoretical analysis on the performance of Algorithm 1. We firstly propose our main theorem that upper bounds the cumulative regret.

Theorem D.1 (Regret). *Let $n_0 = 6(\log_{4/3} T + \check{C}_K)$ where $\check{C}_K := \log_{4/3} p_{\max} + 1$ and $\delta_K = \sqrt{2\log\frac{48(2mn+1)T}{\epsilon}} \cdot \max\{p_{\max}, \gamma_{\max}\} I_{\max}$. Algorithm 1 guarantees an $\tilde{O}(\sqrt{Tmn} + mn)$ regret with probability at least $1-\epsilon$. Here $\tilde{O}(\cdot)$ omits the dependence on $\log\frac{1}{\epsilon}$ and $\log T$.*

This regret rate is near-optimal with respect to T, as it matches the information-theoretic lower bound of $\Omega(\sqrt{T})$ (see Broder and Rusmevichientong, 2012, Theorem 3.1), which describes a special case as $m = n = 1$ and $\gamma_1 = C_{1,1} = 0$ in our setting.

D.1 Sub-Regret and Confidence Bound of $\mathcal{A}_K$

In this subsection, we present two lemmas that show the convergence of each agent $\mathcal{A}_K$. Specifically, Lemma D.2 shows the "horizontal convergence" of (inventory, price) decisions towards the local optimal. Lemma D.3 shows the "vertical convergence" of estimation error Δ_K such that we are maintaining and updating a valid lower-confidence bound for the majority of time.

Lemma D.2 (Sub-regret of every $\mathcal{A}_K$). *For agent $\mathcal{A}_K$ (defined as Algorithms 3 to 5) that has been running for T_K time periods so far, the cumulative sub-regret is bounded by:*

$$SReg_K = \sum_{t_K=1}^{T_K} Q(\vec{I}_{t_K}, p_{t_K}) - W(p_K^*) = \tilde{O}(\sqrt{T_K}). \tag{10}$$

The proof of Lemma D.2 is relegated to Appendix F.4.

Lemma D.3 (Validity of Δ_K). *For any agent $\mathcal{A}_K$ that has been running for T_K time periods with $T_K \geq 6(\log_{4/3} T + \check{C}_K)$, we have $\Delta_K = \tilde{O}(\frac{1}{\sqrt{T_K}})$.*

The proof of Lemma D.3 is in Appendix F.5. From Lemma D.3, we directly get the following corollary.

Corollary D.4. *After at most $N_0 := 6(mn+1)(\log_{4/3} T + \check{C}_K)$ time periods, there does* not *exist any $K \in [mn] \cup \{0\}$ such that $\Delta_K = +\infty$.*

Furthermore, due to the piecewise convexity of $W(p)$ and the convergence rate of Δ_K, combining with Corollary D.4, we have the following lemma.

Lemma D.5. *At any time $t > N_0 := 6(mn+1)(\log_{4/3} T + \check{C}_K)$, we have*

$$LCB_K \leq W(p_K^*) \text{ and } LCB_K \geq W(p_K^*) - 35\Delta_K - \frac{2L_W}{T}. \tag{11}$$

Here L_W is the Lipschitz coefficient of $W(p)$.

The proof details of Lemma D.5 is displayed in Appendix F.6. With the help of all the lemmas above, we are now ready to provide an upper bound on the total regret.

$$\begin{aligned}
Reg_T :=& \sum_{K=0}^{mn} SReg_K + \sum_{t=1}^{T} W(p_{K_t}^*) - W(p^*) \\
\leq& \sum_{K=0}^{mn} \tilde{O}(\sqrt{T_K}) + \sum_{t=N_0}^{T} (35\Delta_K + \frac{2L_W}{T}) + 6(mn+1)(\log_{4/3} T + \check{C}_K) \cdot a_{\max} \cdot p_{\max} \\
=& \tilde{O}(\sqrt{mn \cdot T} + \sum_{K=0}^{mn} \sum_{t_K=1}^{T_K} \frac{1}{\sqrt{t_K}} + mn) \\
=& \tilde{O}(\sqrt{Tmn} + mn).
\end{aligned} \tag{12}$$

E More Discussions

Here we discuss potential extensions of our work, aiming for connections between existing results and future research.

Pricing and Inventory Co-Decisions. The incorporation of inventory constraints into dynamic pricing problems began with the work of Besbes and Zeevi (2009) which assumed a fixed initial stock, and decisions of replenishment were later allowed in Chen et al. (2019). More recent studies, including Chen et al. (2020); Keskin et al. (2022), assumed perishable goods and took inventory costs into account. The stream of work by Chen et al. (2021a, 2023) further assumed the inventory-censoring effect on demands. However, none of them consider the heterogeneity of supply, nor the impact of prices on the allocation process. In our work, we not only model the inventory cost of each warehouse individually, but also depict the unit supplying cost from Warehouse i to Consumer j as a unique coefficient $C_{i,j}$.

Online Convex Optimization (OCO). OCO models a scenario in which decisions are made iteratively, facing a series of convex loss functions, with the objective of minimizing cumulative regret over time (Shalev-Shwartz et al., 2012) or within certain budgets (Jenatton et al., 2016). In our work, we adopt zeroth-order methods in Agarwal et al. (2011) when iterating within each local convexity interval. For a detailed review of the classic and contemporary results on OCO, we kindly refer the readers to Hazan (2016).

Generalization to non-linear demands. We assume the demand $\vec{D}$ is a linear function of price p, which is a widely-used assumption (see LaFrance, 1985). Meanwhile, we still want to generalize our methodologies to a broader family of non-linear demands. Notice that the second-stage allocation problem defined by Eq. (5) does not involve the formulation of demand w.r.t. p. Therefore, we may still divide the price space into $[C_{i_K,j_K}, C_{i_{K+1},j_{K+1}}]$ intervals, and run an individual online optimization agent within each interval. With a similar analysis, we can achieve an $\tilde{O}(T^\alpha (mn)^{1-\alpha})$ regret, where $\alpha \geq 1/2$ is dependent on the demand family we assume. On the other hand, by selecting $m = n = 1$ and $C_{i,j} = 0$, we may have a lower bound at $\Omega(T^\alpha)$.

Generalization to censored demands. In this work, we consider a warehouse-retailer setting where the demand orders are realized and informed to the suppliers *before* they are served. However, there exists another supply-demand relationship, such as groceries and wholesales, where the realized demands are revealed only *after* the resources are delivered from the supply side to the demand side as a preparation. In that case, we should estimate the prospective demand and carefully balance the allocation among individuals in each side respectively, which goes much beyond a straightforward LP scheme as we solve Eq. (5). Besides, the realized demand might be *censored* when supply shortage occurs, making the problem more challenging. Therefore, we expect future investigations toward that new problem.

Pricing and service fairness. Our model maintains fairness in the pricing process by offering the same price to all consumers. However, while the greedy policy for resource allocation is reasonable, widely adopted, and analytically optimal, it leads to differentiated service levels among consumers. We anticipate future research focused on ensuring fairness in service levels during resource allocation.

F Proof Details

F.1 Proof of Lemma B.1

Notice that Eq. (5) defines a linear programming which contains a matrix variable with constraints on the (weighted) sum of each row and each column. Therefore, we may prove a generalized version of Lemma B.1, which is defined as follows.

Lemma F.1. *Given parameters $c \in \mathbb{R}^s, A \in \mathbb{R}_+^{m\times s}, B \in \mathbb{R}_+^{n\times s}$, define the following optimization problem*

$$\begin{aligned} g(\vec{I},\vec{D}) := \min_{x\in\mathbb{R}^s} \quad & c^\top x \\ \text{s.t.} \quad & x \succeq 0 \\ & Ax \preceq \vec{I} \\ & Bx \preceq \vec{D}. \end{aligned} \tag{13}$$

It holds that $g(\vec{I},\vec{D})$ is convex w.r.t. $[\vec{I};\vec{D}]$.

Proof of Lemma F.1. Consider the Lagrangian of Eq. (13):

$$L(x,\mu,\lambda,\eta;\vec{I},\vec{D}) := c^\top x - \mu^\top x + \lambda^\top (Ax-\vec{I}) + \eta^\top (Bx-\vec{D}) = (c-\mu+A^\top\lambda+B^\top\eta)^\top x - \lambda^\top\vec{I} - \eta^\top\vec{D}. \tag{14}$$

Therefore, we have

$$\begin{aligned} g(\vec{I},\vec{D}) :=& \min_x \max_{\mu,\lambda,\eta\succeq 0} L(x,\mu,\lambda,\eta;\vec{I},\vec{D}) \\ =& \max_{\mu,\lambda,\eta\succeq 0} \min_x L(x,\mu,\lambda,\eta;\vec{I},\vec{D}) \\ =& \max_{\mu,\lambda,\eta\succeq 0} \min_x (c-\mu+A^\top\lambda+B^\top\eta)^\top x - \lambda^\top\vec{I} - \eta^\top\vec{D} \\ =& \max_{\mu,\lambda,\eta\succeq 0} -\lambda^\top\vec{I} - \eta^\top\vec{D} \\ & \text{s.t.} \quad c-\mu+A^\top\lambda+B^\top\eta = 0. \end{aligned} \tag{15}$$

Here the second line is due to the strong duality of linear programming. Since the last line indicates that $g(\vec{I},\vec{D})$ can be represented as the piecewise max of linear functions (which is convex), we know that $g(\vec{I},\vec{D})$ is also convex w.r.t. $[\vec{I};\vec{D}]$ jointly. ■

F.2 Proof of Lemma B.3

We denote $\tilde{C}_{i,j}(p) := -p + C_{i,j}$. For each fixed $k \in [mn]$, for any $p \in (C_{i_k,j_k}, C_{i_{k+1},j_{k+1}})$, the sign of any $\tilde{C}_{i,j}(p)$ is fixed: For every $(i',j') \succeq (i_{k+1},j_{k+1})$, we know that $\tilde{C}_{i',j'}(p) = -p + C_{i',j'} \geq -p + C_{i_{k+1},j_{k+1}} > -p + p = 0$. Therefore, in the optimal solution of $g(\vec{I},p,\vec{D}) = \min_{\mathbf{X}} \sum_{i=1}^m \sum_{j=1}^n \tilde{C}_{i,j} X_{i,j}$, we can assign $X_{i',j'} = 0$ for any $(i',j') \succeq (i_{k+1},j_{k+1})$ without loss of generality (as its coefficient is positive). Given this, there exists $A \in \mathbb{R}^{k\times k}, B \in \mathbb{R}^{k\times k}, \vec{C} \in \mathbb{R}^k$ such that the optimization problem defined as Eq. (5) for $p \in (C_{i_k,j_k}, C_{i_{k+1},j_{k+1}})$ can be generalized to the following linear programming:

$$\begin{aligned} g(\vec{I},p,\vec{D}) := & \min_{x\in\mathbb{R}^k} (-p\vec{\mathbb{1}} + \vec{C})^\top x \\ & \text{s.t. } x \succeq 0 \\ & Ax \preceq \vec{I} \\ & Bx \preceq \vec{D}. \end{aligned} \tag{16}$$

Without loss of generality, in the following part of this proof of Lemma B.3, we show that $g(\vec{I}, p, \vec{D})$ defined in Eq. (16) is convex for $p \in (C_{i_k,j_k}, C_{i_{k+1},j_{k+1}})$. The Lagrangian of this new $g(\vec{I}, p, \vec{D})$ is

$$\begin{aligned} &L(x, \mu, \lambda, \eta; \vec{I}, \vec{D}) \\ :=&(-p \cdot \vec{\mathbb{1}} + \vec{C})^\top x - \mu^\top x + \lambda^\top (Ax - \vec{I}) + \eta^\top (Bx - \vec{D}) \\ =&(-p \cdot \vec{\mathbb{1}} + \vec{C} - \mu + A^\top \lambda + B^\top \eta)^\top x - \lambda^\top \vec{I} - \eta^\top \vec{D}. \end{aligned} \tag{17}$$

Since linear programming has strong duality, we further have

$$\begin{aligned} g(\vec{I}, p, \vec{D}) &= \max_{\mu,\lambda,\eta} \min_{x} L(x, \mu, \lambda, \eta, \vec{I}, \vec{D}) \\ &= \max_{\mu,\lambda,\eta \geq 0} -\lambda^\top \vec{I} - \eta^\top \vec{D} \\ &\qquad \text{s.t.} \; -p \cdot \vec{\mathbb{1}} + \vec{C} - \mu + A^\top \lambda + B^\top \eta = 0 \\ &= \max_{\lambda,\eta \geq 0} -\lambda^\top \vec{I} - \eta^\top \vec{D} \\ &\qquad \text{s.t.} \; -p \cdot \vec{\mathbb{1}} + \vec{C} + A^\top \lambda + B^\top \eta \geq 0. \end{aligned} \tag{18}$$

As a consequence, the definition of $Q_t(\vec{I}, p)$ is now generalized to

$$\begin{aligned} Q_t(\vec{I}, p) &:= \langle \vec{\gamma}, \vec{I} \rangle + g(\vec{I}, p, \vec{a} - \vec{b}p + \vec{N}_t) \\ &= \langle \vec{\gamma}, \vec{I} \rangle + \max_{\lambda,\eta \geq 0} -\lambda^\top \vec{I} - \eta^\top (\vec{a} - \vec{b}p + \vec{N}_t) \\ &\qquad \text{s.t.} \; p \cdot \vec{\mathbb{1}} \leq A^\top \lambda + B^\top \eta + \vec{C}. \end{aligned} \tag{19}$$

Denote

$$S(p) := \{(\lambda, \eta) | p \cdot \vec{\mathbb{1}} \leq A^\top \lambda + B^\top \eta + \vec{C}, \lambda \geq 0, \eta \geq 0\}. \tag{20}$$

Notice that for any p_1, p_2 s.t. $C_{i_k,j_k} \leq p_1 < p_2 \leq C_{i_{k+1},j_{k+1}}$, we have $S(p_1) \supseteq S(p_2)$, indicating that $S(p)$ is a monotonically shrinking convex set as p increases. Given the definition of $S(p)$ in Eq. (20), the definition of $Q_t(\vec{I}, p)$ is equivalent to

$$Q_t(\vec{I}, p) := \max_{(\lambda,\eta) \in S(p)} -\lambda^\top \vec{I} - \eta^\top (\vec{a} - \vec{b}p + \vec{N}_t) \tag{21}$$

In the following, we prove a more generalized lemma, from which we can immediately derive the convexity of Eq. (21).

Lemma F.2. *Consider a family of functions $\mathcal{F} := \{f(x; \theta)\}$, where $f(x; \theta)$ is Lipschitz and convex on $x \in [0, 1]$ and is parametrized by θ. $S(x)$ is a convex set that is monotonically not expanding w.r.t. x (i.e., $S(x_1) \supseteq S(x_2)$ if $0 \leq x_1 < x_2 \leq 1$). If $Q(x) := \max_{\theta \in S(x)} f(x; \theta)$ is Lipschitz, then $Q(x)$ is convex on $[0, 1]$.*

Proof of Lemma F.2. Denote the epigraph of a function $g : \mathbb{R} \to \mathbb{R}$ as

$$epi_{[l,r]} \, g(\cdot) := \{(x, t) | x \in [l, r], t \geq g(x)\}. \tag{22}$$

Also, for set $S(x), x \in [0, 1]$, denote

$$S^{-1}(\theta) := \{x \in [0, 1] | \theta \in S(x)\} \tag{23}$$

Since $S(x)$ is monotonically not expanding, we know that $S^{-1}(\theta)$ is continuous. This is because $\theta \in S(x)$ is a sufficient condition of $\theta \in S(y)$ for any $y \in [0, x]$. Given these definitions, we have

$$\begin{aligned} epi_{[0,1]}\, Q(\cdot) =& \{(x,t) | x \in [0,1], t \geq Q(x)\} \\ =& \bigcup_{x\in[0,1]} \{(x,t) | t \geq \max_{\theta \in S(x)} f(x;\theta)\} \\ =& \bigcup_{x\in[0,1]} \{(x,t) | \forall \theta \in S(x), t \geq f(x;\theta)\} \\ =& \bigcup_{x\in[0,1]} \bigcap_{\theta \in S(x)} \{(x,t) | t \geq f(x;\theta)\} \\ =& \bigcap_{\theta\in S(0)} \bigcup_{x \in S^{-1}(\theta)} \{(x,t) | t \geq f(x;\theta)\} \\ =& \bigcap_{\theta\in S(0)} epi_{[0,\sup S^{-1}(\theta)]}\, f(\cdot;\theta). \end{aligned} \tag{24}$$

Since $f(x;\theta)$ is a convex function with respect to x in $[0, 1]$, we know that $epi_{[a,b]} f(\cdot;\theta)$ is a convex domain for any $0 \leq a \leq b \leq 1$. Therefore, their intersections over all $\theta \in S(0)$ is a convex domain. This shows that the epigraph of $Q(x)$ is convex, and therefore $Q(x)$ is convex. ■

With Lemma F.2, we know that $Q_t(\vec{I}, p)$ defined in Eq. (21) is convex with respect to p. This ends the proof of Lemma B.3.

F.3 Proof of Lemma B.4

Denote $\vec{I}^*(p) := \operatorname{argmin}_{\vec{I}} Q(\vec{I}, p)$. In the following, for the clarity of derivatives calculation, we also denote $\vec{I}^* := \vec{I}^*(p)$ for fixed p (i.e. not being differentiated). According to the definition of $W(p)$ given by Eq. (8), we have

$$\begin{aligned} W'(p) =& \frac{d\, Q(\vec{I}^*(p), p)}{d\, p} \\ =& \frac{\partial Q(\vec{I}^*, p)}{\partial p} + \frac{\partial Q(\vec{I}, p)}{\partial I}\Big|_{\vec{I}=\vec{I}^*(p)} \cdot \frac{d\, \vec{I}^*(p)}{d\, p}. \end{aligned} \tag{25}$$

According to the definition of $\vec{I}^*(p)$, we have $\frac{\partial Q(\vec{I},p)}{\partial \vec{I}} = 0$ at $\vec{I} = \vec{I}^*(p)$. Now we show that $\frac{\partial Q(\vec{I}^*(p),p)}{\partial p}$ is a monotonically increasing function of p. According to the definition of Q, we have

$$\begin{aligned} & Q(\vec{I}^*(p), p) \\ =& \mathbb{E}[Q_t(\vec{I}^*(p), p)] \\ =& \mathbb{E}[g(\vec{I}^*(p), p, \vec{a} - \vec{b} \cdot p + \vec{N}_t) + \langle \vec{\gamma}, \vec{I}^*(p) \rangle] \\ =& \int_{N_1} \int_{N_2} \cdots \int_{N_n} g(\vec{I}^*(p), p, \vec{a} - \vec{b} \cdot p + [N_1, N_2, \ldots, N_n]^\top) \cdot \rho_{\vec{N}_t}(N_1, N_2, \ldots, N_n) dN_1 dN_2 \ldots dN_n + \langle \vec{\gamma}, \vec{I}^*(p) \rangle \\ =& \lim_{\Delta \to 0^+} \lim_{C_j = C \to +\infty} \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \cdots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \Big(g(\vec{I}^*(p), p, \vec{a} - \vec{b} \cdot p + [-C_1 + l_1\Delta, -C_2 + l_2\Delta, \ldots, -C_n + l_n\Delta]^\top) + \langle \vec{\gamma}, \vec{I}^*(p) \rangle \Big) \\ & \cdot \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta])]. \end{aligned} \tag{26}$$

Here $\rho_{\vec{N}_t}(N_1, N_2, \ldots, N_n)$ is the probabilistic density function (PDF) of the noise $\vec{N}_t$ at $\vec{N}_t = [N_1, N_2, \ldots, N_n]$. Also, denote

$$\check{Q}_\Delta(\vec{I}^*, p) := \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \cdots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \Big(g(\vec{I}^*, p, \vec{a} - \vec{b} \cdot p + [-C_1 + l_1\Delta, -C_2 + l_2\Delta, \ldots, -C_n + l_n\Delta]^\top) + \langle \vec{\gamma}, \vec{I}^* \rangle \Big) \cdot \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta])]. \tag{27}$$

Since the limit operation preserves convexity, the convexity of $\check{Q}_\Delta(\vec{I}^*(p), p)$ (with respect to p) is a sufficient condition of the convexity of $Q(\vec{I}^*(p), p)$. To prove the convexity of $\check{Q}_\Delta(\vec{I}^*(p), p)$, we firstly show that there are *at most* $E := (\frac{2C}{\Delta})^n \cdot 2^{\binom{mn+m+n}{mn}}$ singularities in each $[C_{i_K, j_K}, C_{i_{K+1}, j_{K+1}}]$ interval such that $\check{Q}_\Delta(\vec{I}^*(p), p)$ is non-smooth. In fact, since $g(\vec{I}, p, \vec{D})$ is the value of a linear program (LP), the solution should be located at the vertex of its feasible space. According to Eq. (5), there are $(mn + m + n)$ one-dimensional linear constraints in the $\mathbb{R}^{m \times n}$ space. As a consequence, the feasible domain of $\mathbf{X} \in \mathbb{R}_+^{m \times n}$ has at most $\binom{mn+m+n}{mn}$ vertices. For the LP defined by $g(\vec{I}^*, p, \vec{a} - \vec{b} \cdot p + [-C_1 + l_1\Delta, -C_2 + l_2\Delta, \ldots, -C_n + l_n\Delta]^\top)$, the value of $\sum_{i=1}^m \sum_{j=1}^n (-p + C_{i,j}) X_{i,j}$ at each vertex of the feasible domain of $\mathbf{X}$ is smooth with respect to p. Denote those vertices as $V_1, V_2, \ldots, V_G$ as $\binom{G := mn+m+n}{mn}$, and the value of $\sum_{i=1}^m \sum_{j=1}^n (-p + C_{i,j}) X_{i,j}$ at each vertex V_l as $h_l(p)$. Again, the solution of an LP should be located at the vertex, and hence we have $g(\vec{I}^*, p, \vec{a} - \vec{b} \cdot p + [-C_1 + l_1\Delta, -C_2 + l_2\Delta, \ldots, -C_n + l_n\Delta]^\top) = \min_{l \in \{1,2,\ldots,G\}} h_l(p)$. Since $\{h_l(p)\}_{l=1}^G$ has at most 2^G intersections, we know that $\min_{l \in \{1,2,\ldots,G\}}$ has at most 2^G non-smooth singularities. According to Eq. (27), $\check{Q}_\Delta(\vec{I}^*, p)$ is a summation of $(\frac{2C}{\Delta})^n$ different $g(\cdot, \cdot, \cdot)$ functions. Therefore, it has at most $(\frac{2C}{\Delta})^n \cdot 2^{\binom{mn+m+n}{mn}}$ non-smooth singularities.

Without loss of generality, denote them as

$$C_{i_K, j_K} \leq P_1 < P_2 < \ldots < P_E \leq C_{i_{K+1}, j_{K+1}}.$$

Also denote $P_0 := C_{i_K, j_K}$ and $P_{E+1} := C_{i_{K+1}, j_{K+1}}$. Now we propose another two lemmas.

Lemma F.3. *For $p \in (P_e, P_{e+1}), e = 0, 1, 2, \ldots, E$, we show that $\frac{\partial \check{Q}_\Delta(\vec{I}^*, p)}{\partial p}$ is monotonically increasing on p.*

Proof of Lemma F.3. Notice that

$$\begin{aligned} \frac{\partial}{\partial p} \check{Q}_\Delta(\vec{I}^*, p) &= \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \cdots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \frac{\partial g(\vec{I}^*, p, \vec{a} - \vec{b}p + \vec{N}_t)}{\partial p} \Big|_{\check{\vec{N}}_t} \cdot \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta])] \\ &= \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \cdots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \frac{\partial Q_t(\vec{I}^*, p)}{\partial p} \Big|_{\check{\vec{N}}_t} \cdot \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta])]. \end{aligned} \tag{28}$$

Here $\check{\vec{N}}_t := [-C_1 + l_1\Delta, -C_2 + l_2\Delta, \ldots, -C_n + l_n\Delta]^\top$. Now we consider the monotonicity of $\frac{\partial Q_t(\vec{I}^*, p)}{\partial p}$ on each (P_e, P_{e+1}) interval. Since there exist no singularities in this interval, we know that

$Q_t(\vec{I}^*, p) \in \mathbb{C}^2$ in this range, and therefore we have

$$\begin{aligned}\frac{d}{d\,p}\frac{\partial Q_t(\vec{I}^*, p)}{\partial p} :=& \langle \frac{\partial Q_t(\vec{I}^*, p)}{\partial \vec{I} \partial p}, \frac{d\,\vec{I}^*(p)}{d\,p}\rangle + \frac{\partial^2 Q_t(\vec{I}^*, p)}{\partial p^2} \\ =& \langle \frac{\partial}{\partial p}\frac{\partial Q_t(\vec{I}^*, p)}{\partial \vec{I}}|_{\vec{I}=\vec{I}^*(p)}, \frac{d\,\vec{I}^*(p)}{d\,p}\rangle + \frac{\partial^2 Q_t(\vec{I}^*, p)}{\partial p^2} \\ =& \langle \vec{0}, \frac{d\,\vec{I}^*(p)}{d\,p}\rangle + \frac{\partial^2 Q_t(\vec{I}^*, p)}{\partial p^2} \\ =& \frac{\partial^2 Q_t(\vec{I}^*, p)}{\partial p^2} \geq 0.\end{aligned} \tag{29}$$

Here the second line that we swap the sequence of derivatives is due to the smoothness within the (P_e, P_{e+1}) smooth interval, and the last line is from Lemma B.3 which shows the marginal convexity of $Q_t(\vec{I}, p)$ w.r.t. p. Therefore, we have proved the lemma. ∎

Lemma F.4. *At each P_e for $e = 0, 1, 2, \ldots, E$, we have $W'(P_e^-) \leq W'(P_e^+)$.*

Proof of Lemma F.4. We firstly consider $W'(P_e^-)$. According to the proof of Lemma F.3, we have

$$\begin{aligned}W'(P_e^-) =& \lim_{p \to P_e^-} W'(p) \\ =& \lim_{p \to P_e^-} \frac{\partial}{\partial p} \lim_{\Delta \to 0^+} \lim_{C \to +\infty} \check{Q}_\Delta(\vec{I}^*, p) \\ =& \lim_{p \to P_e^-} \lim_{\Delta \to 0^+} \lim_{C \to +\infty} \frac{\partial}{\partial p} \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \ldots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} Q_t(\vec{I}^*, p)|_{\check{\vec{N}}_t} \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta] \\ =& \lim_{\Delta \to 0^+} \lim_{C \to +\infty} \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \ldots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta] \cdot \lim_{p \to P_e^-} \frac{\partial Q_t(\vec{I}^*, p)}{\partial p} \\ \leq& \lim_{\Delta \to 0^+} \lim_{C \to +\infty} \sum_{l_1=1}^{\frac{2C_1}{\Delta}} \sum_{l_2=1}^{\frac{2C_2}{\Delta}} \ldots \sum_{l_n=1}^{\frac{2C_n}{\Delta}} \Pr[\bigcap_{j=1}^{n} (\vec{N}_t(j) \in [-C_j + (l_j - 1)\Delta, -C_j + l_j\Delta] \cdot \lim_{p \to P_e^+} \frac{\partial Q_t(\vec{I}^*, p)}{\partial p} \\ =& \lim_{p \to P_e^+} \frac{\partial}{\partial p} \lim_{\Delta \to 0^+} \lim_{C \to +\infty} \check{Q}_\Delta(\vec{I}^*, p) \\ =& \lim_{p \to P_e^+} W'(p) \\ =& W'(p_e^+).\end{aligned} \tag{30}$$

Here the fourth and the sixth lines are due to the Moore-Osgood theorem of exchanging limits, as $Q_t(\vec{I}^*, p) \in \mathbb{C}^2$ when $p \in (P_{e-1}, P_e)$ and $p \in (P_e, P_{e+1})$ respectively. The fifth line comes from Lemma B.3: as $Q_t(\vec{I}, p)$ is convex w.r.t. p, the left derivatives of $\frac{\partial Q_t(\vec{I}, p)}{\partial p}$ should not exceed its right derivatives at any point p. ∎

Applying Lemma F.3 on Eq. (25), we know that $W(p)$ is convex within each smooth interval (P_e, P_{e+1}). Also, from Lemma F.4, we know that $W(p)$ is convex at any singularity P_e as its left derivatives does not exceed its right derivatives. Combining those two properties, we know that $Q(\vec{I}^*, p)$ is convex w.r.t. p. This ends the proof.

F.4 Proof of Lemma D.2

The main idea of this proof originates from OCO with zeroth-order (bandit) feedback, as is displayed in Agarwal et al. (2011). Specifically, we conduct the proof in the following steps:

(a) When an agent $\mathcal{A}_K$ is in Stage 1, Epoch τ and Sub-Epoch s, then we sequentially show that

(i) The aggregated function $Q_{K,\tau,s}(\vec{I},p)$ is concentrated to $Q(\vec{I},p)$ for the three proposed $p=\hat{p}_\tau \in \{a_{K,\tau}, b_{K,\tau}, c_{K,\tau}\}$ and for any $\vec{I}$, with $O(1/\sqrt{n_s})$ error.

(ii) The $\hat{Q}_{K,\tau,s,p}$, which takes the empirical optimal inventory decision, is concentrated to $Q(\vec{I}^*(p),p) = W(p)$ at those proposed prices $p=\hat{p}_\tau \in \{a_{K,\tau}, b_{K,\tau}, c_{K,\tau}\}$, with $O(1/\sqrt{n_s})$ error.

(iii) According to the convexity of $W(p)$, we upper bound the sub-regret per round $W(p) - W(p_K^*)$ by $O(1/\sqrt{n_s})$, where p_K^* is the local optimal price.

(iv) We show that the total number of epochs in Stage 1 is $O(\log T)$. According to the doubling lengths of n_s, the total sub-regret of $\mathcal{A}_K$ is $\tilde{O}(\sqrt{T_K})$ by the time when $\mathcal{A}$ has proposed T_K pairs of decisions $(\vec{I}_t, p_t)$.

(b) When an agent $\mathcal{A}_K$ reaches Stage 2 or 3, we know that the search space $[L_K, U_K]$ is smaller than $1/T$. Given that the Q_t functions (and therefore $W(p)$) are Lipschitz, we may upper bound the sub-regret per step as $O(1/T)$ and the total sub-regret as $O(1)$.

Before we get to proof details, we propose a lemma that generally holds for convex functions.

Lemma F.5. *Suppose $f : [a,b] \to \mathbb{R}$ is a L-Lipschitz convex function. Denote $f(x^*) := \min_{x\in[a,b]} f(x), x_1 = \frac{3a+b}{4}, x_2 = \frac{a+b}{2}, x_3 = \frac{a+3b}{4}$. Assume there exists some fixed constants A and $\Delta > 0$ such that $f(x_i) \in [A-\Delta, A+\Delta], i=1,2,3,$ then we have*

$$\max\{f(x_1), f(x_2), f(x_3)\} - f(x^*) \le 4\Delta. \tag{31}$$

Please kindly find the proof of Lemma F.5 in Appendix F.7.

Now we return to the main proof. We firstly propose the following concentration lemma

Lemma F.6. *For Agent $\mathcal{A}_K$ running in Epoch τ Sub-Epoch s, and $\forall \vec{I} \in \mathbb{R}_+$, symbolic variable $\hat{p}_\tau \in \{a_{K,\tau}, b_{K,\tau}, c_{K,\tau}\}$, we have*

$$|Q_{K,\tau,s}(\vec{I},\hat{p}_\tau) - Q(\vec{I},\hat{p}_\tau)| \le \frac{C_c}{2\sqrt{n_s}} \tag{32}$$

with probability $\Pr \ge 1-\hat{\epsilon}$. Here $C_c := \sqrt{2\log\frac{2}{\hat{\epsilon}}} \cdot Q_{\max}$ and $Q_{\max} := \max\{p_{\max}, \gamma_{\max}\} I_{\max}$.

We will specify the value of $\hat{\epsilon}$ as a function of ϵ by the end of this proof. We defer the proof of Lemma F.6 to Appendix F.8. From Lemma F.6, we may get the following corollary.

Corollary F.7. *For Agent $\mathcal{A}_K$ running in Epoch τ Sub-Epoch s, and symbolic variable $\hat{p}_\tau \in \{a_{K,\tau}, b_{K,\tau}, c_{K,\tau}\}$ we have*

$$\begin{aligned} |\hat{Q}_{K,\tau,s,\hat{p}_\tau} - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)| &\le \frac{C_c}{2\sqrt{n_s}}, \text{ and} \\ 0 \le Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau) - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) &\le \frac{C_c}{\sqrt{n_s}} \end{aligned} \tag{33}$$

with probability $\Pr \ge 1-4T\hat{\epsilon}$.

Proof of Corollary F.7. For each fixed tuple $(K,\tau,s,\hat{p}_\tau)$, according to Lemma F.6, we have

$$\begin{aligned} \hat{Q}_{K,\tau,s,\hat{p}_\tau} - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) =& Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau) - Q_{K,\tau,s}(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) + Q_{K,\tau,s}(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) \\ \le& 0 + \frac{C_c}{2\sqrt{n_s}} = \frac{C_c}{2\sqrt{n_s}}. \\ \hat{Q}_{K,\tau,s,\hat{p}_\tau} - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) =& Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau) - Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau) + Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau) - Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau) \\ \ge& -\frac{C_c}{2\sqrt{n_s}} + 0 = -\frac{C_c}{2\sqrt{n_s}} \end{aligned} \tag{34}$$

with probability $\Pr \ge 1-2\hat{\epsilon}$. Here the first inequality comes from the optimality of $I_{K,\tau,s}(\hat{p}_\tau)$ over $Q_{K,\tau,s}(\cdot,\hat{p}_\tau)$ as well as the concentration of $Q_{K,\tau,s}(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)$ towards $Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)$. The second inequality comes from the concentration of $Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)$ towards $Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)$ as well as the optimality of $\vec{I}^*(\hat{p}_\tau)$ over $Q(\cdot,\hat{p}_\tau)$.

Also, we have

$$
\begin{aligned}
&Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)-Q(I^*(\hat{p}_\tau),\hat{p}_\tau)\\
=&Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)-Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)\\
&+Q_{K,\tau,s}(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)-Q_{K,\tau,s}(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)\\
&+Q_{K,\tau,s}(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)-Q(\vec{I}^*(\hat{p}_\tau),\hat{p}_\tau)\\
\leq&\frac{C_c}{2\sqrt{n_s}}+0+\frac{C_c}{2\sqrt{n_s}}\\
=&\frac{C_c}{\sqrt{n_s}}
\end{aligned}
\tag{35}
$$

with probability $\Pr\geq 1-2\hat{\epsilon}$. Here the third line comes from the concentrations (the first and the third term) as well as the optimality of $I_{K,\tau,s}(\hat{p}_\tau)$ over $Q_{K,\tau,s}(\cdot,\hat{p}_\tau)$. Besides, the other side that $Q(I_{K,\tau,s}(\hat{p}_\tau),\hat{p}_\tau)-Q(I^*(\hat{p}_\tau),\hat{p}_\tau)$ is due to the optimality of $\vec{I}^*(\hat{p}_\tau)$ over $Q(\cdot,\hat{p}_\tau)$.

Since the combination of $(K,\tau,s,\hat{p}_\tau)$ is unique, and the total number of combinations is exactly T, we apply the union bound of probability and get that Eq. (33) holds for all $(K,\tau,s,\hat{p}_\tau)$ tuples with probability $\Pr\geq 1-4T\hat{\epsilon}$. ■

Combining Corollary F.7 with Lemma F.5, we have the following corollary.

Corollary F.8. *Define a $flag$ as shown in Algorithm 3, and define $p^*_K := \operatorname{argmin}_{p\in[C_{i_K,j_K},C_{i_{K+1},j_{K+1}}]} W(p)$. When $flag == 0$ by the end of Sub-Epoch s of Epoch τ, we have*

$$
|W(\hat{p}_\tau)-W(p^*_K)|\leq 16\cdot\frac{C_c}{\sqrt{n_s}}
\tag{36}
$$

holds for $\hat{p}_\tau=a_{K,\tau},b_{K,\tau},c_{K,\tau}$ with probability $\Pr\geq 1-24T\hat{\epsilon}$.

Proof of Corollary F.8. When $flag == 0$, according to Algorithm 3, we know that

$$
\begin{aligned}
|\hat{Q}_{K,\tau,s,a_{K,\tau}}-\hat{Q}_{K,\tau,s,b_{K,\tau}}|&\leq 4\varDelta_{K,\tau,s}\\
\hat{Q}_{K,\tau,s,a_{K,\tau}}-\hat{Q}_{K,\tau,s,c_{K,\tau}}&\leq 4\varDelta_{K,\tau,s}\\
\hat{Q}_{K,\tau,s,b_{K,\tau}}-\hat{Q}_{K,\tau,s,c_{K,\tau}}&\leq 4\varDelta_{K,\tau,s}.
\end{aligned}
\tag{37}
$$

Also, according to the convexity of $W(p)$ in $[C_{i_K,j_K},C_{i_{K+1},j_{K+1}}]$, we know that $W(c_{K,\tau})\leq\frac{W(a_{K,\tau})+W(b_{K,\tau})}{2}\leq\max\{W(a_{K,\tau}),W(b_{K,\tau})\}$. Without loss of generality, assume $W(a_{K,\tau})\geq W(b_{K,\tau})$, and then we have

$$
\begin{aligned}
\hat{Q}_{K,\tau,s,c_{K,\tau}}-\hat{Q}_{K,\tau,s,a_{K,\tau}}=&Q(I_{K,\tau,s}(c_{K,\tau}),c_{K,\tau})-Q(\vec{I}^*(c_{K,\tau}),c_{K,\tau})\\
&+W(c_{K,\tau})-W(a_{K,\tau})\\
&+Q(\vec{I}^*(a_{K,\tau}),a_{K,\tau})-Q(I_{K,\tau,s}(a_{K,\tau}),a_{K,\tau})\\
\leq&\frac{C_c}{\sqrt{n_s}}+0+\frac{C_c}{\sqrt{n_s}}\\
=&\frac{2C_c}{\sqrt{n_s}}\leq 4\varDelta_{K,\tau,s}.
\end{aligned}
\tag{38}
$$

Therefore we know that $|\hat{Q}_{K,\tau,s,b_{K,\tau}} - \hat{Q}_{K,\tau,s,c_{K,\tau}}| \leq 4\Delta_{K,\tau,s}$. Combining Corollary F.7, we have

$$\begin{aligned}
|W(a_{K,\tau}) - W(b_{K,\tau})| \leq & |Q(\vec{I}^*(a_{K,\tau}), a_{K,\tau}) - Q(I_{K,\tau,s}(a_{K,\tau}), a_{K,\tau})| \\
& + |Q(I_{K,\tau,s}(a_{K,\tau}), a_{K,\tau}) - Q(I_{K,\tau,s}(b_{K,\tau}), b_{K,\tau})| \\
& + |Q(I_{K,\tau,s}(b_{K,\tau}), b_{K,\tau}) - Q(\vec{I}^*(b_{K,\tau}), b_{K\tau})| \\
\leq & \frac{C_c}{\sqrt{n_s}} + 4\Delta_{K,\tau,s} + \frac{C_c}{\sqrt{n_s}} \\
\leq & \frac{4C_c}{\sqrt{n_s}}.
\end{aligned} \tag{39}$$

And also

$$\begin{aligned}
|W(a_{K,\tau}) - W(c_{K,\tau})| \leq & |Q(\vec{I}^*(a_{K,\tau}), a_{K,\tau}) - Q(I_{K,\tau,s}(a_{K,\tau}), a_{K,\tau})| \\
& + |Q(I_{K,\tau,s}(a_{K,\tau}), a_{K,\tau}) - Q(I_{K,\tau,s}(b_{K,\tau}), c_{K,\tau})| \\
& + |Q(I_{K,\tau,s}(c_{K,\tau}), b_{K,\tau}) - Q(\vec{I}^*(b_{K,\tau}), c_{K\tau})| \\
\leq & \frac{C_c}{\sqrt{n_s}} + 4\Delta_{K,\tau,s} + \frac{C_c}{\sqrt{n_s}} \\
\leq & \frac{4C_c}{\sqrt{n_s}}.
\end{aligned} \tag{40}$$

By applying Lemma F.5 with $\Delta = \frac{4C_c}{\sqrt{n_s}}$, we show that the lemma holds with $\Pr \geq 1 - 24T\hat{\epsilon}$ (since we have used Corollary F.7 for 6 times). ■

Finally, we show an upper bound on the total number of epochs in which $\mathcal{A}_K$ is running, and we first denote this number as M_K. In fact, from the design of $Algorithm$ 3, we know that by the end of each epoch, we have $U_{K,\tau+1} - L_{K,\tau+1} = \frac{3}{4}(U_{K,\tau} - L_{K,\tau})$, i.e. the length of search space $[L_{K,\tau}, U_{K,\tau}]$ reduces by $1/4$. Since $L_{K,1} = C_{i_K,j_K}, U_{K,1} = C_{i_{K+1},j_{K+1}}$, we have

$$\begin{aligned}
M_K &\leq \log_{3/4} \frac{U_{K,M_K} - L_{K,M_K}}{U_{K,1} - L_{K,1}} \\
&= \log_{4/3} \frac{U_{K,1} - L_{K,1}}{U_{K,M_K} - L_{K,M_K}} \\
&\leq \log_{4/3} \frac{C_{i_{K+1},j_{K+1}} - C_{i_K,j_K}}{\frac{1}{T} \cdot \frac{3}{4}} \\
&= \log_{4/3}(C_{i_{K+1},j_{K+1}}) + \log_{4/3} T + 1 \\
&\leq \log_{4/3} p_{\max} + \log_{4/3} T + 1.
\end{aligned} \tag{41}$$

Denote $\check{C}_K := \log_{4/3} p_{\max} + 1$, and we have $M_K \leq \log_{4/3} T + \check{C}_K$.

With all the properties above, we may derive the total sub-regret for $\mathcal{A}_K$. Firstly, the cumulative sub-regret in Epoch τ Sub-Epoch s is

$$\begin{aligned}
& SubReg(\mathcal{A}_K, \tau, s) \\
:= & n_s \cdot \sum_{\hat{p}_\tau = a_{K,\tau}, b_{K,\tau}, c_{K,\tau}} Q(\vec{I}_{K,\tau,s-1}(\hat{p}_\tau), \hat{p}_\tau) - Q(\vec{I}^*(p_K^*), p_K^*) \\
\leq & n_s \cdot \sum_{\hat{p}_\tau = a_{K,\tau}, b_{K,\tau}, c_{K,\tau}} |Q(\vec{I}_{K,\tau,s-1}(\hat{p}_\tau), \hat{p}_\tau) - Q(\vec{I}^*(\hat{p}_\tau), \hat{p}_\tau)| + |Q(\vec{I}^*(\hat{p}_\tau), \hat{p}_\tau) - Q(\vec{I}^*(p_K^*), p_K^*)| \\
\leq & n_s \cdot \sum_{\hat{p}_\tau = a_{K,\tau}, b_{K,\tau}, c_{K,\tau}} \frac{C_c}{\sqrt{n_s}} + 20 \cdot \frac{C_c}{\sqrt{n_s}} \\
= & 63 C_c \sqrt{n_s} \\
= & 63 C_c \cdot 2^{s/2}.
\end{aligned} \tag{42}$$

Secondly, denote the number of sub-epochs in Epoch τ as S_τ and the length of Epoch τ as T_τ (therefore we know that $T_\tau = 3 \cdot 2^{S_\tau+1} - 1$), and the cumulative sub-regret in Epoch τ is bounded by

$$\begin{aligned} SubReg(\mathcal{A}_K, \tau) &:= \sum_{s=1}^{S_\tau} SubReg(\mathcal{A}_K, \tau, s) \\ &\le \sum_{s=1}^{S_\tau} 63C_c \cdot 2^{s/2} \\ &\le 63C_c \cdot \frac{1}{\sqrt{2}-1} 2^{\frac{S_\tau+1}{2}} \\ &\le 200C_c \cdot \sqrt{\frac{T_\tau+1}{3}} \\ &\le 200C_c\sqrt{T_\tau}. \end{aligned} \tag{43}$$

Thirdly, we may calculate the total sub-regret of $\mathcal{A}_K$ as

$$\begin{aligned} SReg_K &:= \sum_{\tau=1}^{M_K} SubReg(\mathcal{A}_K, \tau) \\ &\le \sum_{\tau=1}^{M_K} 200C_c\sqrt{T_\tau} \\ &\le 200C_c \cdot \sqrt{\left(\sum_{\tau=1}^{M_K} (\sqrt{T_\tau})^2\right) (\sum_{\tau=1}^{M_K} 1)} \\ &= 200C_c(\sum_{\tau=1}^{M_K} T_\tau)(M_K) \\ &\le 200C_c \cdot \sqrt{T_K \cdot M_K} \\ &\le 200C_c\sqrt{T_K(\log_{4/3} T + \check{C}_K)} \\ &= \tilde{O}(\sqrt{T_K}). \end{aligned} \tag{44}$$

This rate holds with $\Pr \ge 1 - 24T\hat{\epsilon}$ for each $K \in [2mn+1]$. Let $\hat{\epsilon} := \frac{\epsilon}{24\cdot(2mn+1)T}$ so that $C_c = \delta_K = \sqrt{2\log\frac{48(2mn+1)T}{\epsilon}} \cdot \max\{p_{\max}, \gamma_{\max}\} \cdot I_{\max}$, and we complete the proof of Lemma D.2.

F.5 Proof of Lemma D.3

We analyze the behavior of Δ_K by considering the current stage of $\mathcal{A}_K$.

1. If $\mathcal{A}_K$ is currently in Stage 1. Suppose $\mathcal{A}_K$ has played for τ_K epochs. Since each epoch reduces the price interval $[L_{K,\tau}, U_{K,\tau}]$ to its 3/4, we know that $\tau_K \le M_k \le \log_{4/3} T + \check{C}_K$ where $C_K := \log_{4/3} p_{\max} + 1$ (see Eq. (41)).

   According to Pigeon-Hole Theorem, at least the longest epoch $\hat{\tau}$ has been played for $\frac{T_K}{\tau_K} \ge \frac{6(\log_{4/3} T + C_K)}{\log_{4/3} T + C_K} = 6$ times. As a result, at least **two** sub-epochs have been reached in this epoch. Since we update Δ_K by the end of each sub-epoch (except for the last sub-epoch of each epoch), at least one of these two sub-epochs leads to an update on Δ_K. As a result, $\Delta_K < +\infty$ after this update, and after $T_K \ge 6(\log_{4/3} T + C_K)$ time periods.

   Denote the length of this epoch $\hat{\tau}$ as $H_{K,\hat{\tau}}$, and we know that $H_{K,\hat{\tau}} \ge \frac{T_K}{\log_{4/3} T + C_K}$. Also, we denote the length of each sub-epoch of Epoch $\hat{\tau}$ as $H_{K,\hat{\tau},s}$, $s = 1, 2, \ldots, S_{\hat{\tau}}$, where $S_{\hat{\tau}}$ is denoted as the number of sub-epochs of Epoch $\hat{\tau}$. Given those definitions, we have

$$\begin{aligned} H_{K,\hat{\tau},S_{\hat{\tau}}-1} &\geq \sum_{s=1}^{S_{\hat{\tau}}-2} H_{K,\hat{\tau},s} \\ H_{K,\hat{\tau},S_{\hat{\tau}}-1} &\geq \frac{H_{K,\hat{\tau},S_{\hat{\tau}}}}{2}. \end{aligned} \tag{45}$$

As a consequence, we have

$$H_{K,\hat{\tau},S_{\hat{\tau}}-1} \geq \frac{H_{K,\hat{\tau}}}{4} \geq \frac{T_K}{4(\log_{4/3} T + C_K)}. \tag{46}$$

Since we still can update Δ_K by the end of Epoch $\hat{\tau}$ Sub-Epoch $S_{\hat{\tau}} - 1$, we may upper bound Δ_K in the following approach

$$\begin{aligned} \Delta_K =& \frac{\delta_K}{2\sqrt{n_{S_{\hat{\tau}}-1}}} \\ =& \frac{\delta_K}{2 \cdot \sqrt{H_{K,\hat{\tau},S_{\hat{\tau}}-1}/3}} \\ \leq& \frac{\delta_K}{\frac{2}{\sqrt{3}} \cdot \sqrt{\frac{T_K}{4(\log_{4/3} T + C_K)}}} \\ =& \sqrt{\frac{1}{T_K}} \cdot \delta_K \cdot \sqrt{3(\log_{4/3} T + C_K)} \\ =& \tilde{O}(\sqrt{\frac{1}{T_K}}). \end{aligned} \tag{47}$$

2. If $\mathcal{A}_K$ reaches Stage 2. Since we only run Stage 2 for once without stopping, updating $\hat{W}_K, \Delta_K$ or switching agents, we assume that T_K reaches the end of Stage 2 without loss of generality. We firstly upper and lower bound the length of Stage 2. Denote $T_{K,1}$ as the time periods that $\mathcal{A}_K$ spent in Stage 1, and $T_{K,2} := T_K - T_{K,1}$ as the time periods that $\mathcal{A}_K$ has spent in Stage 2 so far. Recall that the purpose of conducting Stage 2 is to guarantee a Δ_K that is comparable to $\sqrt{\frac{1}{T_K}}$, and at the end of Stage 2 we reduce Δ_K to its half comparing to the one we have by the end of Stage 1 (if not $+\infty$). Therefore, we have

$$2T_{K,1} \geq 2H_{K,\hat{\tau}} \geq N_{K,2} \geq H_{K,\hat{\tau}} \geq \frac{T_{K,1}}{\log_{4/3} T + C_K}. \tag{48}$$

Here the first inequality represents that $\mathcal{A}_K$ runs $T_{K,1}$ time periods in Stage 1, including $H_{K,\hat{\tau}}$ time periods in Stage 1 Epoch $\hat{\tau}$. The second and third inequalities hold because we get comparable Δ_K in Stage 1 and in Stage 2, and the best Δ_K we got in Stage 1 is on the longest sub-epoch, which is $\hat{\tau}$. The last inequality is from the proof shown in Case 1 (when $\mathcal{A}_K$ reaches Stage 1).

Also, since Stage 2 applies a "Doubling Trick", we have

$$\begin{aligned} T_{K,2} \leq& 2^1 + 2^2 + \ldots + N_{K,2} \\ \leq& 2N_{k,2}, \\ \Rightarrow \qquad T_{K,2} \leq& 4T_{K,1} \\ \Rightarrow \qquad T_{K,1} \leq T_K \leq& 5T_{K,1}. \end{aligned} \tag{49}$$

As a result, we have

$$\Delta_K = \frac{\delta_K}{\sqrt{N_{K,2}}} \leq \frac{\delta_K}{\sqrt{\frac{T_{K,1}}{\log_{4/3} T + C_K}}} \leq \frac{\delta_K(\log_{4/3} T + C_K)}{\sqrt{\frac{T_K}{5}}} = \tilde{O}(\frac{1}{\sqrt{T_K}}). \tag{50}$$

3. If $\mathcal{A}_K$ reaches Stage 3. Denote $T_{K,3} := T_K - T_{K,1} - T_{K,2}$ as the time periods that $\mathcal{A}_K$ has spent in Stage 3 so far. According to Algorithm 5, we know that $\Delta_K = \frac{\delta_K}{\sqrt{N_{K,3}}} =$

$\frac{\delta_K}{\sqrt{N_{K,2}+T_{K,3}}} = \tilde{O}(\frac{1}{\sqrt{N_{K,2}+T_{K,3}}})$. Also, since

$$\begin{aligned} N_{K,2}+T_{K,3} \geq & \frac{1}{2}\cdot N_{K,2} + \frac{1}{2}\cdot N_{K,2} + T_{K,3} \\ \geq & \frac{1}{2}\cdot \frac{T_{K,1}}{\log_{4/3} T + C_K} + \frac{1}{2}\cdot\frac{T_{K,2}}{2} + T_{K,3} \\ \geq & \frac{T_{K,1}+T_{K,2}+T_{K,3}}{2(\log_{4/3} T + C_K)} \\ = & \frac{T_K}{2(\log_{4/3} T + C_K)}. \end{aligned} \tag{51}$$

Therefore, we have $\Delta_K \leq \tilde{O}(\frac{1}{\sqrt{\frac{T_K}{2(\log_{4/3} T + C_K)}}}) = \tilde{O}(\frac{1}{\sqrt{T_K}})$. This ends the proof of Lemma D.3.

F.6 Proof of Lemma D.5

We consider each case where $\mathcal{A}_K$ is in Stage 1,2,3, respectively.

1. If $\mathcal{A}_K$ is in Stage 1. When updating Δ_K, we know that $flag == 0$ at Stage 1 Epoch τ Sub-Epoch $s-1$ according to Algorithm 3. Denote $\hat{p} := \mathrm{argmin}_{\hat{p}_\tau \in \{a_{K,\tau}, c_{K,\tau}, b_{K,\tau}\}} \hat{Q}_{K,\tau,s-1,\hat{p}_\tau}$, and we have that

$$\begin{aligned} LCB_K = & \hat{W}_K - 34\Delta_K \\ = & \hat{Q}_{K,\tau,s-1}(I_{K,\tau,s-1}(\hat{p}),\hat{p}) - 34\cdot\frac{\delta_K}{2\sqrt{n_{s-1}}} \\ = & \hat{Q}_{K,\tau,s-1}(I_{K,\tau,s-1}(\hat{p}),\hat{p}) - W(\hat{p}) + W(\hat{p}) - 17\cdot\frac{\delta_K}{\sqrt{n_{s-1}}} \\ \leq & \frac{C_c}{\sqrt{n_{s-1}}} - \frac{\delta_K}{\sqrt{n_{s-1}}} + W(\hat{p}) - 16\cdot\frac{\delta_K}{\sqrt{n_{s-1}}} \\ \leq & 0 + W(p_K^*). \end{aligned} \tag{52}$$

Here the fourth line comes from Corollary F.7 and the last line is an application of $C_c = \delta_K$ and Corollary F.8.

On the other hand, the lower bound LCB_K is not too faraway from $W(p_K^*)$ since we have:

$$\begin{aligned} & LCB_K \\ = & \min\{Q_{K,\tau,s-1,a_{K,\tau}}, Q_{K,\tau,s-1,c_{K,\tau}}, Q_{K,\tau,s-1,b_{K,\tau}}\} - 34\Delta_K \\ = & \min\{Q_{K,\tau,s-1,a_{K,\tau}} - \frac{C_c}{2\sqrt{n_{s-1}}}, Q_{K,\tau,s-1,c_{K,\tau}} - \frac{C_c}{2\sqrt{n_{s-1}}}, Q_{K,\tau,s-1,b_{K,\tau}} - \frac{C_c}{2\sqrt{n_{s-1}}}\} - 33\cdot\frac{C_c}{2\sqrt{n_{s-1}}} \\ \geq & \min\{W(a_{K,\tau}), W(b_{K,\tau}), W(c_{K,\tau})\} - 33\cdot\frac{C_c}{2\sqrt{n_{s-1}}} \\ \geq & W(p_K^*) - 33\Delta_K. \end{aligned} \tag{53}$$

2. If $\mathcal{A}_K$ is in Stage 2 and Stage 3, we may consider them altogether as the only update of LCB as well as Δ_K occurs by the end of Stage 2, which is also the 0-th time period of Stage 3. In

this case, we have

$$\begin{aligned}
&\hat{W}_K - W(p_K^*)\\
&= (\hat{W}_K - \hat{Q}_K^*) + (Q_{K,r_K}(\vec{I}_K^*, \hat{p}_K^*) - Q_{K,r_K}(\vec{I}^*(\hat{p}_K^*), \hat{p}_K^*))\\
&\quad + (Q_{K,r_K}(\vec{I}^*(\hat{p}_K^*), \hat{p}_K^*) - Q(\vec{I}^*(\hat{p}_K^*), \hat{p}_K^*)) + (W(\hat{p}_K^*) - W(p_K^*))\\
&\le 0 + 0 + \frac{\delta_K}{2\sqrt{N_{K,2} + T_{K,3}}} + \frac{L_W}{T}\\
&\le \Delta_K + \frac{L_W}{T}.\\
&W(p_K^*) - \hat{W}_K\\
&= (W(p_K^*) - W(\hat{p}_K^*)) + (W(\hat{p}_K^*) - Q(\vec{I}_K^*, \hat{p}_K^*)) + (Q(\vec{I}_K^*, \hat{p}_K^*) - Q_{K,r_K}(\vec{I}_K^*, \hat{p}_K^*)) + (\hat{Q}_K^* - \hat{W}_K)\\
&\le 0 + 0 + \frac{\delta_K}{2\sqrt{N_{K,2} + T_{K,3}}} + \frac{L_W}{T}\\
&\le \Delta_K + \frac{L_W}{T}.
\end{aligned} \tag{54}$$

Since $LCB_K = \hat{W}_K - 34\Delta_K - \frac{L_W}{T}$, we have

$$\begin{aligned}
LCB_K &\le W(p_K^*) + \Delta_K + \frac{L_W}{T} - 34\Delta_K - \frac{L_W}{T}\\
&\le W(p_K^*).\\
LCB_K &\ge W(p_K^*) - \Delta_K - \frac{L_W}{T} - 34\Delta_K - \frac{L_W}{T}\\
&\ge W(p_K^*) - 35\Delta_K - \frac{2L_W}{T}.
\end{aligned} \tag{55}$$

Combining the two cases listed above, we have proved this lemma.

F.7 Proof of Lemma F.5

Proof. Denote $f_1 := f(x_1), f_2 := f(x_2), f_3 := f(x_3)$. Then we prove this lemma by cases where x^* locates.

1. When $x^* \in [a, x_1]$, we denote $\epsilon := \frac{x_2 - x_1}{x_2 - x^*}$. We know that $\epsilon \in [\frac{1}{2}, 1]$, and then we have $\epsilon f(x^*) + (1-\epsilon) f(x_2) \ge f(\epsilon x^* + (1-\epsilon) x_2) = f(x_1)$ due to the convexity of $f(x)$. As a result, we have:

$$\begin{aligned}
f(x^*) \ge& \frac{f_1 - (1-\epsilon) f_2}{\epsilon}\\
=& f_2 + \frac{f_1 - f_2}{\epsilon}\\
=& f_2 - \frac{f_2 - f_1}{\epsilon}.
\end{aligned} \tag{56}$$

If $f_1 \ge f_2$, then we have $f(x^*) \ge f_2 \ge A - \Delta = A + \Delta - 2\Delta \ge \max\{f_1, f_2, f_3\} - 2\Delta$. Otherwise $f_1 < f_2$, then we have

$$\begin{aligned}
f(x^*) &\ge f_2 - \frac{f_2 - f_1}{\epsilon}\\
&\ge f_2 - \frac{f_2 - f_1}{\frac{1}{2}}\\
&= 2f_1 - f_2\\
&\ge 2(A - \Delta) - (A + \Delta)\\
&= A - 3\Delta\\
&= A + \Delta - 4\Delta\\
&\ge \max\{f_1, f_2, f_3\} - 4\Delta.
\end{aligned}$$

2. When $x^* \in (x_1, x_2]$, let $\epsilon = \frac{x_3 - x_2}{x_3 - x^*}$, and the proof goes the same way as in (1).

3. When $x^* \in (x_2, x_3]$, we let $\epsilon = \frac{x_2 - x_1}{x^* - x_1}$ and we know that $\epsilon \in [\frac{1}{2}, 1]$. Since $x_2 = \epsilon \cdot x^* + (1 - \epsilon)x_1$, we have $\epsilon f(x^*) + (1-\epsilon) f(x_1) \geq f(x_2)$ according to the convexity of $f(x)$. Therefore, we have:

$$\begin{aligned} f(x^*) \geq & \frac{f_2 - (1-\epsilon) f_1}{\epsilon} \\ = & f_1 + \frac{f_2 - f_1}{\epsilon} \\ = & f_1 - \frac{f_1 - f_2}{\epsilon}. \end{aligned} \tag{57}$$

   If $f_1 \leq f_2$, then we have $f(x^*) \geq f_1 \geq A - \Delta = A + \Delta - 2\Delta \geq \max\{f_1, f_2, f_3\} - 2\Delta$. Otherwise $f_1 > f_2$, then we have

$$\begin{aligned} f(x^*) &\geq f_1 - \frac{f_1 - f_2}{\epsilon} \\ &\geq f_1 - \frac{f_1 - f_2}{\frac{1}{2}} \\ &= 2f_2 - f_1 \\ &\geq 2(A - \Delta) - (A + \Delta) \\ &= A - 3\Delta \\ &= A + \Delta - 4\Delta \\ &\geq \max\{f_1, f_2, f_3\} - 4\Delta. \end{aligned}$$

4. When $x^*(x_3, b]$, let $\epsilon = \frac{x_3 - x_2}{x^* - x_2}$, and the proof goes the same way as (3).

■

F.8 Proof of Lemma F.6

Proof. Notice that

$$-p_{\max} I_{\max} \leq -p \cdot \|\vec{I}\|_1 \leq Q_t(\vec{I}, p) \leq \langle \vec{\gamma}, \vec{I} \rangle \leq \gamma_{\max} I_{\max} \tag{58}$$

Denote $Q_{\max} := \max\{p_{\max}, \gamma_{\max}\} I_{\max}$. By applying Hoeffding's Inequality to $\forall \vec{I}, \hat{p}_\tau \in \{a_{K,\tau}, b_{K,\tau}, c_{K,\tau}\}$, we have

$$\begin{aligned} & \Pr[|Q_{K,\tau,s}(\vec{I}, \hat{p}_\tau) - Q(\vec{I}, \hat{p}_\tau)| \geq x] < 2\exp\{-\frac{2x^2 n_s}{Q_{\max}^2}\} \\ \Rightarrow \quad & |Q_{K,\tau,s}(\vec{I}, \hat{p}_\tau) - Q(\vec{I}, \hat{p}_\tau)| \leq \frac{Q_{\max}\sqrt{2\log\frac{2}{\hat{\epsilon}}}}{2\sqrt{n_s}} \text{ with } \Pr \geq 1 - \hat{\epsilon}. \end{aligned} \tag{59}$$

Let $C_c = Q_{\max}\sqrt{2\log\frac{2}{\hat{\epsilon}}}$ and this completes the proof. ■

G Conclusion

In this paper, we study an online learning problem under the framework of pricing and allocation, where we make joint pricing and inventory decisions and allocate supplies to fulfill demands over time. To solve this non-convex problem, we propose a hierarchical algorithm which incorporates an LCB meta-algorithm over multiple local OCO agents. Our analysis shows that it guarantees an $\tilde{O}(\sqrt{Tmn} + mn)$ regret, which is optimal with respect to T as it matches the existing lower bound. Our work sheds light on the cross-disciplinary research of machine learning and operations research, especially from an online perspective.