Learning in Continuous State-Space MDPs for Network Inventory Management

Hansheng Jiang
University of Toronto

Shunan Jiang
University of California, Berkeley

Zuo-Jun Shen
University of Hong Kong

Abstract

We consider online learning in infinite-horizon, average-cost Markov Decision Processes (MDPs) with multi-dimensional, continuous state spaces and censored feedback. Our model setting, motivated by network inventory management applications such as vehicle sharing, is characterized by complex, correlated state transitions and the absence of value function convexity, rendering standard analytical techniques for both MDPs and inventory control inapplicable. Our primary contribution is a integrated framework establishing and leveraging the Lipschitz property of the long-run average cost function. This insight allows us to analyze the problem through the lens of Lipschitz bandits, for which we design a provably efficient online learning algorithm that learns a near-optimal policy from censored demand data. We derive a high-probability regret bound of $O(T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}})$, where n is the network size through customized concentration inequalities for cumulative costs in MDPs with state-dependent transitions. Furthermore, we devise a matching lower bound for this learning problem, which captures the inherent dimensionality challenge.

1 INTRODUCTION

Modern urban mobility platforms, such as vehicle-sharing systems, represent a grand challenge for sequential decision-making. The platforms must dynamically reposition a fleet of vehicles across a network to meet unknown, spatially-correlated demand, balancing high operational costs against the opportunity cost of lost demand. This operational problem is a canonical instance of an infinite-horizon, average-cost Markov Decision Process (MDP) with a multi-dimensional, continuous state space.

Solving such MDPs in an online learning setting, where demand distributions are unknown, presents a confluence of several distinct theoretical challenges. First, the continuous, multi-dimensional nature of the state space precludes direct tabular methods and requiring structural assumptions to enable generalization. Second, the learner receives only censored feedback; true customer demand is only observed up to available inventory. This information loss is a well-known impediment in inventory theory and online learning, breaking the full-information assumption of many bandit and MDP analyses. Third, actions (repositioning) have correlated, system-wide consequences, inducing complex temporal correlations in the state transitions. This violates the i.i.d. or simple mixing assumptions that underpin standard concentration inequalities used in regret analysis. Finally, and perhaps most critically, the value function in this networked setting lacks the convexity that is the cornerstone of analysis in the vast majority of the single-product inventory control literature. This structural failure renders the powerful tools of stochastic convex optimization inapplicable and necessitates a fundamentally new approach.

To overcome these obstacles, we develop an integrated algorithmic and theoretical framework. Our contributions are:

1. **Provably Efficient Algorithm:** Our `LipBR` algorithm learns a near-optimal inventory repositioning policy without prior knowledge of demand distributions. We prove a high-probability regret bound of $O(T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}})$, where n is the number of locations (the dimensionality of the decision). This result generalizes the classic $\widetilde{O}(\sqrt{T})$ regret rates from single-location inventory control (Huh et al., 2009) to multi-location networks.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

2. **Lipschitz Bandits-Based Approach:** We leverage the Lipschitz continuity of the long-run average cost function in the policy space. In contrast to prior inventory learning methods that rely on convex cost structures (Agrawal and Jia, 2022; Yuan et al., 2021), we show that the average-cost objective in our networked MDP is Lipschitz continuous with respect to changes in the base-stock policy. This insight allows us to cast the problem into a Lipschitz multi-armed bandit framework (Kleinberg et al., 2008). By doing so, we circumvent the need for value-function convexity and instead exploit smoothness: our analysis "zooms in" on a near-optimal policy by gradually refining a cover of the continuous policy space. This technique could be of independent interest for other continuous-state MDP problems.

3. **New Theoretical Analysis:** We derive several technical results to enable the above analysis. Most notably, we establish customized concentration inequalities for cumulative costs in an MDP with state-dependent, correlated transitions. These bounds extend classical martingale concentration results to our dependent-cost setting, ensuring that the observed performance of a policy concentrates around its expected long-run cost. We also provide a tight characterization of the covering number of the policy space, which quantifies the exploration complexity in our Lipschitz bandit approach. These analytical tools are not only crucial for our regret bounds but also enrich the theoretical toolkit for analyzing other correlated continuous state-space MDPs.

4. **Optimality and Lower Bound:** We complement our algorithmic results with a matching lower bound. In particular, we prove that any learning algorithm in our setting incurs regret at least on the order of $\Omega(T^{\frac{n}{n+1}})$, even when compared against the best base-stock policy in hindsight. This lower bound, which we derive via a novel information theoretic argument shows that our algorithm's regret rate is order-optimal. It also formally captures the intuitive increase in difficulty with higher-dimensional state spaces. Together, our upper and lower bounds pin down the minimax regret scaling for the addressed problem.

Overall, our work is the first to achieve provably sublinear regret rate in a multi-location inventory control problem without assuming convexity or full feedback. The framework and insights developed here bridge ideas from stochastic inventory theory, Lipschitz bandits, and continuous state-space MDPs, and pave the way for tackling other complex spatial inventory control problems under uncertainty.

2 RELATED LITERATURE

Regret Minimization in Continuous-State MDPs. Regret minimization in MDPs with continuous state spaces is a central challenge in modern reinforcement learning theory. While tabular settings are well-understood via algorithms like UCRL2 (Jaksch et al., 2010), the continuous setting requires structural assumptions to generalize from limited experience. A major line of work exploits metric structures in the state-action space, employing adaptive discretization or zooming techniques to manage the exploration-exploitation trade-off (Ortner and Ryabko, 2012; Sinclair et al., 2019; Song and Sun, 2019). These methods typically yield regret bounds scaling with the covering dimension of the space, often taking the form $\tilde{O}(T^{(d+1)/(d+2)})$. Our work aligns with this metric MDP philosophy but diverges in methodology: rather than discretizing the state space, we exploit the Lipschitz property of the average-cost objective function itself. This allows us to apply discretization directly in the policy space, effectively converting the RL problem into a Lipschitz bandit problem.

Lipschitz Bandits and Policy Search. Our algorithmic approach is inspired by the literature on Lipschitz bandits, where the reward function over a continuous action space is assumed to be Lipschitz continuous (Kleinberg et al., 2008). This assumption allows an algorithm to generalize from a finite number of samples to the entire continuum of actions. While classic Lipschitz bandit algorithms provide a powerful framework for stateless optimization, our work extends this paradigm to a full MDP setting. This extension introduces significant new challenges not present in the standard bandit formulation: the costs are state-dependent, the state transitions are correlated over time, and the objective is to optimize a long-run average cost, not a static reward function. Our analysis, therefore, requires establishing the Lipschitz property of the long-run average cost function itself, a nontrivial undertaking. Furthermore, the correlated nature of the state sequence invalidates the use of standard concentration inequalities, necessitating the development of new technical tools tailored to our MDP structure.

By establishing that the long-run average cost is Lipschitz continuous with respect to base-stock levels, we bridge the gap between bandit theory and average-cost RL. This places our work in conversation with policy search methods, which treat the MDP as a black-box function to be optimized (Kar and Singh, 2025). However, unlike standard policy search which often lacks finite-time regret guarantees or assumes convexity, our framework provides rigorous sublinear regret bounds $O(T^{\frac{n}{n+1}})$ specifically tailored to the non-convex, censored, and correlated nature of network inventory dynamics.

Network Inventory Management and Vehicle Sharing We situate our work within inventory management and rental networks. Dynamic inventory control has long been modeled as a MDP in operations research (Scarf, 1960). Benjaafar et al. (2022) study a vehicle-sharing system under a discounted-cost MDP and exploit value-function convexity to design an approximate dynamic programming policy. However, such convexity results typically hold only under restrictive assumptions (e.g., single-product or separable costs) and in discounted settings (Hihat et al., 2023). In contrast, we adopt an average-cost criterion—more suitable for ongoing operations—and we avoid assuming any convex structure that would simplify the problem. The classical literature on transshipment and multi-echelon inventory networks also considers multiple locations, but their inventory dynamics are more restrictive, whereas our model permits arbitrary bidirectional inventory flows.

Our problem is substantially more complex than canonical inventory control (Huh et al., 2009; Agrawal and Jia, 2022): the policy is an n-dimensional allocation (an inventory distribution over a network), and the cost function is nonconvex and coupled across locations. Hence, techniques that succeed for a single product do not directly scale. We instead pioneer a Lipschitz-bandit approach to inventory optimization, which, as discussed, replaces convexity with continuity. To the best of our knowledge, our work is the first to establish sublinear regret for a multi-location inventory system with censored demand and adversarially correlated demand rebalancing. This advances the operations management literature by moving beyond asymptotic guarantees via fluid or mean-field approximations (Akturk et al., 2025) and providing finite-time performance bounds for a practical, high-dimensional inventory control problem.

3 MODEL

We consider a closed inventory network operating over discrete time periods $t = 1, 2, \dots$ (He et al., 2020; Benjaafar et al., 2022). The system manages a fixed total quantity of a divisible resource (e.g., inventory, assets), normalized to 1 for convenience, over n locations. The state of the system at the beginning of period t is the inventory distribution across the locations, represented by a vector $\boldsymbol{x}_t = (x_{t,1}, \dots, x_{t,n})$ in the standard simplex $\Delta_{n-1} := \{\boldsymbol{z} \in \mathbb{R}^n \mid \sum_{i=1}^n z_i = 1, z_i \geq 0\}$. Following the tradition in inventory control (Zipkin, 2008), the vector $\boldsymbol{x}_t$ lies in a continuous space.

3.1 Problem Setup

At the start of each period t, after observing the current inventory state $\boldsymbol{x}_t$, a decision maker chooses a target inventory level $\boldsymbol{y}_t \in \Delta_{n-1}$ to reposition the network. Following this decision, stochastic demand $\boldsymbol{d}_t = \{d_{t,i}\}_{i\in[n]}$ occurs at each location. For each location i, only the fulfilled, potentially censored, demand $\min(y_{t,i}, d_{t,i})$ is observed, while the remainder $(d_{t,i} - y_{t,i})^+$ is lost.

This redistribution of inventory, through customers' renting and returning activities, is modeled by a stochastic routing matrix $\boldsymbol{P}_t = (P_{t,ij})_{n\times n}$, where $P_{t,ij}$ is the proportion of fulfilled demand from location i that is transferred to location j. The network is closed, meaning $\sum_{j=1}^n P_{t,ij} = 1$ for all $i \in [n]$. The demand $\boldsymbol{d}_t$ and routing $\boldsymbol{P}_t$ are random and may be correlated. The process $\{(\boldsymbol{d}_t, \boldsymbol{P}_t)\}_{t\geq 1}$ is a sequence of independently and identically distributed (i.i.d.) random variables. The uncensored demand is bounded, i.e., there exists a constant U such that $d_{t,i} < U$ almost surely for all $t \geq 1, i \in [n]$. The inventory transition is given by

$$\boldsymbol{x}_{t+1} = (\boldsymbol{y}_t - \boldsymbol{d}_t)^+ + \boldsymbol{P}_t^\top \min(\boldsymbol{y}_t, \boldsymbol{d}_t). \tag{1}$$

The repositioning cost, $M(\boldsymbol{y}_t - \boldsymbol{x}_t)$, is the minimum cost to move inventory from the current distribution $\boldsymbol{x}_t$ to the target $\boldsymbol{y}_t$, which is computed by the following minimum-cost network flow problem:

$$M(\boldsymbol{y}_t - \boldsymbol{x}_t) = \min \sum_{i=1}^n \sum_{j=1}^n c_{ij}\xi_{ij} \tag{2}$$

$$\text{s.t.} \sum_{i=1}^n \xi_{ij} - \sum_{k=1}^n \xi_{jk} = y_{t,j} - x_{t,j}, \forall j \tag{3}$$

$$\xi_{ij} \geq 0, \forall i, j, \tag{4}$$

where ξ_{ij} is the amount of inventory moved from i to j, and $c_{ij} > 0$ is the unit transportation cost. This linear program is always feasible since $\sum_j (y_{t,j} - x_{t,j}) = 0$. The lost sales cost arises from unmet demand and is given by:

$$L(\boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) = \sum_{i=1}^n \sum_{j=1}^n l_{ij} \cdot P_{t,ij}(d_{t,i} - y_{t,i})^+, \tag{5}$$

where $l_{ij} > 0$ is the unit cost for lost demand that would have moved from location i to j. The total cost for period t is the sum of these components:

$$C(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) = M(\boldsymbol{y}_t - \boldsymbol{x}_t) + L(\boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t). \tag{6}$$

While we assume linear costs, our analysis can be extended to more general convex cost functions.

MDP Formulation The network inventory management problem can be viewed as an infinite-horizon MDP.

1. State: Current inventory distribution $\boldsymbol{x}_t \in \Delta_{n-1}$.
2. Action: Repositioning to the target inventory level $\boldsymbol{y}_t \in \Delta_{n-1}$.
3. Transition: The probabilistic state transition $\mathbb{P}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{y}_t)$ is implicitly defined by (1) and the distribution $\boldsymbol{\mu}$ of $(\boldsymbol{d}_t, \boldsymbol{P}_t)$.
4. Single-Period Cost: $C(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t)$.

A policy π is a mapping from states to actions, $\boldsymbol{y}_t = \pi(\boldsymbol{x}_t)$. Our goal is to find an optimal policy π^* that minimizes the long-run average cost:

$$v^{\pi}(\boldsymbol{x}) := \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi}\left[C(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) \mid \boldsymbol{x}_1 = \boldsymbol{x}\right].$$

We adopt the average-cost criterion, common in inventory theory, for several reasons. First, in systems with frequent decisions, it often reflects the operational reality better than a discounted criterion, where the choice of a discount factor ρ can be arbitrary. Second, for many such systems, the optimal average cost is independent of the initial state, providing a robust, universal benchmark for performance (Agrawal and Jia, 2022).

3.2 Performance Metric

Benchmark Policy The optimal dynamic repositioning policy for our network inventory MDP is notoriously difficult to compute, often proving intractable for realistically sized problems (Benjaafar et al., 2022). This computational challenge necessitates a simple yet theoretically sound benchmark. We evaluate our algorithm against the best base-stock repositioning policy $\boldsymbol{S}^*$, a natural choice for its practical simplicity and strong theoretical guarantees.

$$\boldsymbol{S}^* \in \operatorname*{argmin}_{\boldsymbol{S}\in\Delta_{n-1}} \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\pi_{\boldsymbol{S}}}\left[C(\boldsymbol{x}_t, \boldsymbol{S}, \boldsymbol{d}_t, \boldsymbol{P}_t)\right],$$

where $\pi_{\boldsymbol{S}}$ refers to the base-stock policy with base-stock level $\boldsymbol{S}$. The soundness of this benchmark policy is multifaceted. First, the best base-stock policy is not merely a heuristic; it is proven to be asymptotically optimal in regimes of practical importance, namely when the network is large or when lost sales costs dominate repositioning costs (Jiang et al., 2022). This ensures our benchmark is a strong, near-optimal competitor in relevant scenarios. Second, this choice aligns with a significant body of literature in both inventory control and online learning. In inventory theory, simple policies like base-stock are widely studied for their near-optimal performance (Yuan et al., 2021; Gong and Simchi-Levi, 2024). In online learning, this benchmark is a concrete instance of the standard best fixed policy in hindsight, providing a firm theoretical foundation for our regret analysis.

Modified Costs for Censored Feedback The total cost C_t of period t, as defined in (6), is intractable because part of the total cost, the lost sales cost $\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij} \cdot P_{t,ij}(d_{t,i} - y_{t,i})^+$, is unobservable due to unknown lost demand. To address the censoring issue, we introduce the *modified cost* $\widetilde{C}(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t)$ defined as

$$\begin{aligned}&\widetilde{C}(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t)\\ =&C(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) - \sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}P_{t,ij}d_{t,i}.\end{aligned} \tag{7}$$

After this simple transformation, the modified cost $\widetilde{C}_t$ is observable because $\widetilde{C}_t$ can be rewritten as $\widetilde{C}(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) = M(\boldsymbol{y}_t - \boldsymbol{x}_t) - \sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij} \cdot P_{t,ij}\min\{d_{t,i}, y_{t,i}\}$. The expected difference of the cost C_t and the modified cost $\widetilde{C}_t$ is the expectation of $\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}P_{t,ij}d_{t,i}$, which does not depend on any particular repositioning policy. Therefore, replacing the costs by $\widetilde{C}_t$ does not change the difference of average costs between the two policies, and we can use $\widetilde{C}_t$ instead in regret analysis.

Regret Definition Over a time horizon T, an online learning algorithm ALG sequentially decides the target inventory level $\boldsymbol{y}_t$ based on the current state $\boldsymbol{x}_t$ and historical information $\mathcal{F}_{t-1}$ encompassing observations of censored demand and stochastic routing matrix from previous $t-1$ periods. The incurred modified cost at time t is $\widetilde{C}_t^{\mathtt{ALG}}$ at time t.

Given an initial inventory level $\boldsymbol{x}_1$, the goal is to minimize the expected cumulative costs. Following the tradition of the online learning literature, we minimize the regret of the expected cumulative costs compared to that of a benchmark policy, which, in our case, is the best base-stock repositioning policy that minimizes the cumulative costs.

$$\text{Regret}(T) := \mathbb{E}\sum_{t=1}^{T}\left[\widetilde{C}_t^{\mathtt{ALG}}|\boldsymbol{x}_1\right] - \min_{\boldsymbol{S}\in\Delta_{n-1}}\sum_{t=1}^{T}\mathbb{E}\left[\widetilde{C}_t^{\boldsymbol{S}}|\boldsymbol{x}_1\right].$$

4 ALGORITHM DESIGN

Our online network inventory management problem with unknown distribution of $(\boldsymbol{d}, \boldsymbol{P})$ presents a formidable set of simultaneous challenges for regret minimization.

The decision space (the set of base-stock policies) is a high-dimensional continuum, immediately suggesting a connection to continuum-armed bandits. However, unlike standard bandit problems, the cost of selecting an "arm" (a policy $\boldsymbol{S}$) is not stationary; it depends on the current state $\boldsymbol{x}_t$ of the system. This embedding within an MDP means that observations are temporally correlated, violating the i.i.d. assumption that underpins many standard concentration inequalities. Finally, the learning agent receives only censored feedback, as the true lost sales cost is unobservable.

Our approach is to synthesize techniques from Lipschitz bandits and MDP theory to overcome this unique confluence of challenges. We begin by defining a tractable pseudoregret objective based on an observable cost signal in Section 4.1. We then develop customized concentration inequalities tailored for our state-dependent, correlated setting in Section 4.2. Finally, we introduce our algorithm, `LipBR`, which uses a novel epoch-based structure with memory points to bridge the gap between the state-less bandit framework and the dynamic MDP environment.

4.1 Pseudoregret for MDP

We denote the t-th period expected modified cost conditioning on the state $\boldsymbol{x}_t$ under the base-stock repositioning policy $\pi_{\boldsymbol{S}}$ by

$$\begin{aligned}\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) :=& \mathbb{E}\left[\widetilde{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) \,\middle|\, \boldsymbol{x}_t\right] \\ =& \mathbb{E}_{(\boldsymbol{d},\boldsymbol{P})\sim\boldsymbol{\mu}}[\widetilde{C}(\boldsymbol{x},\boldsymbol{S},\boldsymbol{d},\boldsymbol{P}) \mid \boldsymbol{x}=\boldsymbol{x}_t].\end{aligned} \tag{8}$$

Definition 4.1 (Loss and Bias)**.** Given initial state $\boldsymbol{x}_1 = \boldsymbol{x}$ and a base-stock level $\boldsymbol{S} \in \Delta_{n-1}$, we denote the MDP incurred by applying the base-stock repositioning policy $\pi^{\boldsymbol{S}}$ as $\mathcal{M}(\boldsymbol{S}, \boldsymbol{x})$. The loss $\lambda^{\boldsymbol{S}}(\boldsymbol{x})$ and bias $\beta^{\boldsymbol{S}}(\boldsymbol{x})$ of MDP $\mathcal{M}(\boldsymbol{S}, \boldsymbol{x})$ are respectively defined as follows.

$$\lambda^{\boldsymbol{S}}(\boldsymbol{x}) := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) \,\middle|\, \boldsymbol{x}_1 = \boldsymbol{x}\right],$$

$$\beta^{\boldsymbol{S}}(\boldsymbol{x}) := \mathbb{E}\left[\lim_{T\to\infty} \sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) - \lambda^{\boldsymbol{S}}(\boldsymbol{x}_t) \,\middle|\, \boldsymbol{x}_1 = \boldsymbol{x}\right].$$

Remark 4.2. Our definitions of loss and bias are adapted from the finite-space MDP literature (Puterman, 2014, Section 8.2). In general continuous-space MDPs, the limits defining these terms may not exist. However, the compactness of the state and action spaces in our model ensures that the Markov chain induced by any fixed policy possesses the required ergodicity properties for these limits to be well-defined. Since our regret analysis is independent of any specific discretization scheme, we follow the convention in Agrawal and Jia (2022) and proceed without a formal statement of their existence.

Remark 4.3. For a fixed stationary base-stock policy $\pi_{\boldsymbol{S}}$, the induced controlled Markov chain together with the per-period cost defines a Markov reward process; we retain MDP terminology for consistency.

Based on the definition of loss $\lambda^{\boldsymbol{S}}$, we further introduce the pseudoregret defined as follows,

$$\text{PseudoRegret}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{C}_t \middle| \boldsymbol{x}_1\right] - T\lambda^*, \tag{9}$$

where $\lambda^* := \lambda^{\boldsymbol{S}^*}$ and $\boldsymbol{S}^* \in \operatorname{argmin}_{\boldsymbol{S}\in\Delta_{n-1}} \lambda^{\boldsymbol{S}}(\boldsymbol{x})$. From Lemma A.2 and Lemma A.3, $\lambda^{\boldsymbol{S}}(\boldsymbol{x})$ is independent of the initial state $\boldsymbol{x}$. Therefore, $\boldsymbol{S}^*$ and λ^* are well-defined and do not depend on the initial state.

The regret in (3.2) and the pseudoregret in (9) are usually not equal for two reasons: (1) λ^* is the long-term average cost over an infinite time horizon, while the total costs in (3.2) are computed over a finite time horizon of length T. (2) The best base-stock repositioning policy given by $\boldsymbol{S}^* \in \operatorname{argmin}_{\boldsymbol{S}\in\Delta_{n-1}} \lambda^{\boldsymbol{S}}$ does not necessarily maximize the aggregate regret $\text{Regret}(T, \boldsymbol{S})$ over all $\boldsymbol{S}$. However, we can prove that the difference between the regret and the pseudoregret can be effectively controlled using the concentration inequalities introduced next in Section 4.2.

4.2 Key Concentration Inequalities

A major challenge in analyzing the total costs is that the observed costs $\widetilde{C}_t^{\boldsymbol{S}}$ are not i.i.d. due to the state transition dynamics. Standard concentration bounds for UCB-style algorithms, which rely on this assumption, do not apply. Therefore, we derive customized concentration inequalities specifically for the correlated cost sequence generated by our MDP.

In Lemma 4.5, we prove a concentration bound on N consecutive observed costs and the loss of any policy $\boldsymbol{S}$, which will be useful in our algorithm design. The detailed proof of Lemma 4.5 is in Appendix A. As established in Lemma 4.4 (Puterman, 2014, Theorem 8.2.6), we relate the bias and loss in MDP to decompose the difference, and then we use the celebrated Azuma-Hoeffding inequality for martingale difference sequence to get the final bound.

Lemma 4.4. *For any $\boldsymbol{S}, \boldsymbol{x} \in \Delta_{n-1}$, the loss and bias have the following relation:*

$$\lambda^{\boldsymbol{S}}(\boldsymbol{x}) = \mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x},\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}(\boldsymbol{x}')\right] - \beta^{\boldsymbol{S}}(\boldsymbol{x}), \tag{10}$$

where $\mathbb{P}(\cdot|\boldsymbol{x}, \boldsymbol{S})$ is the probability distribution of the *next state given that the previous state and repositioning level is $(\boldsymbol{x}, \boldsymbol{S})$.*

Lemma 4.5. *Given any base-stock repositioning policy $\boldsymbol{S}$, any starting state $\boldsymbol{x}_1$ and a positive integer N, then for any $\epsilon \in (0,1)$, with probability $1-\epsilon$,*

$$\left|\frac{1}{N}\sum_{t=1}^{N}\widetilde{C}_t^{\boldsymbol{S}} - \lambda^{\boldsymbol{S}}(\boldsymbol{x}_1)\right| \le \frac{\max_{i,j}\{c_{ij}\}}{N} + \sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\left[6\max_{i,j} c_{ij} + 2nU\max_{i,j} l_{ij}\right].$$

In the process of proving Lemma 4.5, we also establish the following two intermediate concentration inequalities, Lemma 4.6 and Lemma 4.7.

Lemma 4.6. *Given a base-stock repositioning level $\boldsymbol{S}$ and a positive integer N, for any $\boldsymbol{x}_1 \in \Delta_{n-1}$, then for any $\epsilon \in (0,1)$, with probability $1-\epsilon/2$,*

$$\left|\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) - \lambda^{\boldsymbol{S}}(\boldsymbol{x}_1)\right| \le \frac{\max_{i,j}\{c_{ij}\}}{N} + 2\sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\max_{i,j}\{c_{ij}\}.$$

Lemma 4.7. *Given an upper bound for the demand level U and a positive integer N, then for any base-stock level $\boldsymbol{S}$, with probability $1-\epsilon/2$,*

$$\left|\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) - \frac{1}{N}\sum_{t=1}^{N}\widetilde{C}_t^{\boldsymbol{S}}\right| \le \sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\left[4\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}\right].$$

The concentration results also imply the following bound on regret vs. pseudoregret.

Corollary 4.8. *With probability $1-T^{-2}$, $|\mathrm{Regret}(T) - \mathrm{PseudoRegret}(T)| \le O(\sqrt{T}\log T)$.*

This $O(\sqrt{T})$ gap is dominated by the main $\widetilde{O}(T^{n/(n+1)})$ bound for $n > 1$, so we design the algorithm around the pseudoregret with its tractable benchmark $T\lambda^*$.

4.3 LipBR Algorithm

4.3.1 Algorithm Description

In Algorithm 1, we define the mean reward to be the negative of the mean cost (see line 10). In this way, we keep the notation similar to the literature. The identified arm with the highest upper confidence bound in line 2 is essentially the arm with the lowest lower confidence bound in terms of the mean cost. At line 11 of Algorithm 1, we update the memory point $\boldsymbol{m}_k$ for arm k to be the inventory level at the end of this epoch (or the beginning of the next time stamp).

Two key ingredients of Algorithm 1 are policy space discretization by covering, and the design of memory points, which we explain in detail in Section 4.3.2 and Section 4.3.3, respectively.

Algorithm 1 `LipBR`: Lipschitz Bandits-based Repositioning Algorithm

```
1: for epoch $i = 1, 2, \ldots,$ do
2:     Identify an arm $k^{(i)}$ with the highest upper confidence bound $k \in \mathrm{argmax}_j\, \mathrm{UB}_j$
3:     for iteration $j = 1, 2, \ldots, N_{k^{(i)}}$ do
4:         if $t > T$ then
5:             End
6:         end if
7:         Reposition inventory level to $\boldsymbol{S}_{k^{(i)}}$, observe the censored demand and calculate the pseudo modified cost $\widetilde{C}_t'$ using equation (11)
8:         Update: $t \leftarrow t + 1$
9:     end for
10:    Update: upper confidence bound $\mathrm{UB}_{k^{(i)}} = \bar{\mu}_{k^{(i)}} + H\sqrt{\frac{\log(T)}{\tau_{k^{(i)}}}}$, where $\tau_{k^{(i)}} = \sum_{l=1}^{t} \mathbb{1}\{a_l = k^{(i)}\}, \bar{\mu}_{k^{(i)}} = -\frac{1}{\tau_{k^{(i)}}}\sum_{l=1}^{t}\mathbb{1}\{a_l = k^{(i)}\}\widetilde{C}_t'$
11:    Update: memory point $\boldsymbol{m}_{k^{(i)}} \leftarrow \boldsymbol{x}_t$; epoch length $N_{k^{(i)}} \leftarrow 2N_{k^{(i)}}$; total epoch count $i \leftarrow i + 1$
12: end for
```

4.3.2 Discretizing Policy Space by Covering

Observing the pseudoregret in (9), one can think of each base-stock repositioning policy $\pi^{\boldsymbol{S}}$ as an arm in bandit problems. Each arm corresponds to a point $\boldsymbol{S}$ in the continuous space Δ_{n-1}, and the optimal arm corresponds to $\boldsymbol{S}^*$. Furthermore, at each time period t, the observed modified cost $\widetilde{C}_t^{\boldsymbol{S}}$ can be viewed as a "noisy" observation on the cost $\lambda^{\boldsymbol{S}}$ of arm $\boldsymbol{S}$ in a *continuum-armed bandits* setting.

While prior work in single-location inventory control often relies on the convexity of the cost function to achieve efficient learning (Agrawal and Jia, 2022), this crucial property does not hold in our networked setting. Instead, we establish a weaker but sufficient structural property: the Lipschitz continuity of the long-run average cost function (Lemma 4.10). This motivates our discretization-based approach, which carefully balances the discretization error with the exploration-exploitation trade-off.

For some $\delta \in (0, 1/2)$ that will be specified later, let $\mathcal{A}_\delta = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_K\}$ be a δ-covering of the state space Δ_{n-1} under the metric induced by the ℓ_1 norm

$\|\cdot\|_1$ such that

$$\Delta_{n-1} \subseteq \bigcup_{k=1}^{K} B(\boldsymbol{S}_k, \delta, \|\cdot\|_1),$$

where $B(\boldsymbol{S}_k, \delta, \|\cdot\|_1)$ represents a ball in n-dimensional Euclidean space that is centered at $\boldsymbol{S}_k$ with a radius of δ under the ℓ_1 norm. The δ-covering set $\mathcal{A}_\delta$ discretizes the continuous space Δ_{n-1} of all base-stock levels into K sets contained in some ℓ_1-ball of radius δ. The smallest number of points needed in any covering is known as the covering number $N(\delta, \Delta_{n-1}, \|\cdot\|_1)$. In Lemma 4.9, we show that the cardinality $K := N(\delta, \Delta_{n-1}, \|\cdot\|_1)$ of the smallest δ-covering $\mathcal{A}_\delta$ can be bounded at the scale of $O\left(\delta^{-n+1}\right)$.

Lemma 4.9 (Bound on Covering Number)**.** *The covering number of Δ_{n-1} is at the scale of $\mathcal{O}(\frac{1}{\delta^{n-1}})$ under ℓ_1 norm.*

Furthermore, we establish the following Lipschitz property of the loss function λ.

Lemma 4.10 (Lipschitz Property)**.** *The loss $\lambda^{\boldsymbol{S}}$ is η-Lipschitz in $\boldsymbol{S}$ in the sense that $|\lambda^{\boldsymbol{S}} - \lambda^{\boldsymbol{S}'}| \leq \eta \|\boldsymbol{S} - \boldsymbol{S}'\|_1$ for all $\boldsymbol{S}, \boldsymbol{S}' \in \Delta_{n-1}$, where the Lipschitz constant is $\eta = \max_{ij} l_{ij} + 6 \max_{ij} c_{ij}$.*

Lemma 4.10 implies that the accuracy of the discretization can be effectively controlled by reducing δ. Therefore, by carefully choosing δ, we can balance the regret from exploration and discretization to achieve the desired regret.

4.3.3 Memory Point and Epoch Structure

For any admissible learning algorithm `ALG`, let $a_t \in [|\mathcal{A}_\delta|]$ index the chosen base-stock level $\boldsymbol{S}_{a_t} \in \mathcal{A}_\delta$ at time t for all t. Suppose `ALG` chooses a base-stock level $a_t = k$ at time t and a cost $\widetilde{C}_t^{\boldsymbol{S}}$ is observed. Based on Lemma 4.5, we know that if `ALG` *consecutively* chooses $\boldsymbol{S}_k$ for N times, the average costs will be an accurate estimation of $\lambda^{\boldsymbol{S}}$ when N is sufficiently large.

However, the problem with a dynamic learning algorithm is that it needs to balance exploration-exploitation and therefore the applied policy cannot be fixed all the time. More concretely, for any policy $\pi^{\boldsymbol{S}}$, suppose t_f is the first time algorithm `ALG` chooses $\pi^{\boldsymbol{S}}$ and t_s is the second time `ALG` chooses $\pi^{\boldsymbol{S}}$, then $\boldsymbol{x}_{t_s}$ is not equal to $\boldsymbol{x}_{t_f+1}$ and thus the concentration bound Lemma 4.5 does not apply because Lemma 4.5 requires N consecutive observed costs.

To overcome this, we introduce a novel epoch-based learning scheme with memory points. This mechanism allows us to "stitch together" cost observations from non-consecutive time steps, creating a simulated sequence of costs that appears to come from a single, uninterrupted run of a fixed policy. This construction is crucial for adapting the concentration bounds from Section 4.2 to the dynamic, explore-exploit setting of the algorithm.

Specifically, we introduce a memory point $\boldsymbol{m}_k$ for each base-stock level $\boldsymbol{S}_k$, which is used to save the state of the system during the last time the policy $\pi^{\boldsymbol{S}_k}$ is used. Suppose at t_s, the policy is newly switched to $\pi^{\boldsymbol{S}_k}$ and $t_f := \max\{t \mid a_t = k, t < t_s\}$ is the last time $\boldsymbol{S}_k$ is chosen by the algorithm; then, we define the memory point as $\boldsymbol{m}_k := \boldsymbol{x}_{t_f+1}$.. We call each time segment that one policy is consecutively used as one epoch, and the memory point is updated at the end of each epoch. The length of each epoch is designed to double every time an arm is being re-pulled in order to establish appropriately tight upper confidence bounds.

On top of memory points, we define the following pseudo modified cost in the `LipBR` algorithm to simulate the cost if the policy were played consecutively. For $t = 1, 2, \ldots,$

$$\widetilde{C}'_t = \begin{cases} \widetilde{L}(\boldsymbol{S}_{a_t}, \boldsymbol{d}_t, \boldsymbol{P}_t) + M(\boldsymbol{S}_{a_t} - \boldsymbol{m}_{a_t}) & \text{if } a_t \neq a_{t-1}; \\ \widetilde{C}_t & \text{otherwise,} \end{cases} \tag{11}$$

where $\boldsymbol{m}_{a_t} = \boldsymbol{x}_{t_0+1}$ and $t_0 := \max\{\tau \mid a_\tau = a_t, \tau < t\}$, and $\widetilde{L}$ is the modified lost sales cost $-\sum_{i=1}^n \sum_{j=1}^n l_{ij} \cdot P_{t,ij} \min\{d_{t,i}, y_{t,i}\}$. When $a_t \neq a_{t-1}$, $M(\boldsymbol{S}_{a_t} - \boldsymbol{m}_{a_t})$ is used instead of $M(\boldsymbol{S}_{a_t} - \boldsymbol{x}_t)$ when calculating the repositioning cost. Based on the sub-additivity of repositioning costs, we have the following relationship between the pseudo modified cost $\widetilde{C}'_t$ and the original modified cost $\widetilde{C}_t$,

$$\begin{aligned} \widetilde{C}_t =& \widetilde{L}(\boldsymbol{S}_{a_t}, \boldsymbol{d}_t, \boldsymbol{P}_t) + M(\boldsymbol{S}_{a_t} - \boldsymbol{m}_{a_t}) \\ &+ M(\boldsymbol{S}_{a_t} - \boldsymbol{x}_t) - M(\boldsymbol{S}_{a_t} - \boldsymbol{m}_{a_t}) \\ \leq& \widetilde{C}'_t + M(\boldsymbol{m}_{a_t} - \boldsymbol{x}_t). \end{aligned} \tag{12}$$

The inequality in (12) implies that the original modified cost $\widetilde{C}_t$ can be effectively bounded by the pseudo modified cost $\widetilde{C}'_t$ up to a repositioning cost from $\boldsymbol{x}_t$ the memory point $\boldsymbol{m}_{a_t}$.

5 REGRET ANALYSIS

5.1 Regret Upper Bound

We now present the main regret bound for our `LipBR` algorithm. The theorem establishes a sublinear regret of $\widetilde{O}\left(T^{n/(n+1)}\right)$, where n is the dimension of the policy space (i.e., the number of locations).

This rate is characteristic of online learning problems in high-dimensional continuous spaces under a Lipschitz assumption, where regret is expected to depend polynomially on the dimension. Our regret rate recovers the classic $\tilde{O}(\sqrt{T})$ scaling for a single location ($n = 1$) and yields a new n-dependent rate $O(T^{n/(n+)}(\log T)^{1/(n+1)})$ for multi-location networks ($n \geq 2$).

Theorem 5.1. *Suppose Algorithm 1 is run with $\delta = (\log T/T)^{1/(n+1)}, H = 2(6\max_{i,j} c_{ij} + 2nU \max_{i,j} l_{ij})$, then the pseudoregret and regret of Algorithm 1 are both upper-bounded by $C_1(\max_{i,j} c_{ij} + nU \max_{i,j} l_{ij})T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}}$, where $C_1 > 0$ is a universal constant independent of the model parameters.*

Proof Sketch of Theorem 5.1. The pseudoregret can be decomposed as follows,

$$\text{PseudoRegret}(T) = \sum_{t=1}^{T} \mathbb{E}[\widetilde{C}_t(\boldsymbol{x}_t)] - T\lambda^*$$
$$\leq \underbrace{\sum_{t=1}^{T} \mathbb{E}[\widetilde{C}'_t(\boldsymbol{x}_t)] - T\lambda^*}_{\text{Regret Part (I)}} + \underbrace{\sum_{t=2}^{T} \mathbb{1}\{a_{l-1} \neq a_l\} M(\boldsymbol{m}_{a_l} - \boldsymbol{x}_t)}_{\text{Regret Part (II)}}.$$

Part (I) captures the exploration-exploitation tradeoff. After accounting for the discretization error, its analysis follows the principles of UCB algorithms for K-armed bandits. A key distinction from standard bandit analysis is the need for our custom concentration inequalities (Section 4.2) to handle the correlated nature of the costs drawn from the MDP. This part contributes a regret of $O(\sqrt{KT\log T})$.

Part (II) is the cumulative "cost of stitching", which represents the overhead incurred by our memory point mechanism, which links non-consecutive episodes of the same policy to create a valid learning signal. We bound this term by showing that the number of policy switches (i.e., new epochs) is at most $K\log(T/K)$ for each arm.

Combining three parts together, the pseudoregret can be bounded by

$$O(\sqrt{KT\log T}) + K\delta + O(K\log(T/K)).$$

We then plug in the value $\delta = (\log T/T)^{1/(n+1)}$ into the above inequality, and then a bound $O\left(T^{n/(n+1)}(\log T)^{1/(n+1)}\right)$ is acquired.

Lastly, Corollary 4.8 states that the difference between the Regret(T) and PseudoRegret(T) is bounded $O(\sqrt{T\log T})$, which is dominated by the pseudoregret $O\left(T^{n/(n+1)}(\log T)^{1/(n+1)}\right)$. This concludes the proof sketch. □

Remark 5.2. If the time horizon T is unknown, we can run `LipBR` in phases of lengths 2^m and restart the algorithm at the start of each phase with parameters tuned to the phase length. This adds at most a logarithmic factor to the regret bound.

Although the regret bound in Theorem 5.1 is asymptotic in T, the leading constant in the LipBR upper bound is explicit: $\max_{i,j} c_{ij} + nU\max_{i,j} l_{ij}$. This highlights the tradeoff between repositioning and lost-sales costs: (i) when repositioning dominates, the constant is essentially driven by $\max_{i,j} c_{ij}$ and is largely insensitive to U. (ii) when lost sales dominate, the constant grows linearly in n through $U\max_{i,j} l_{ij}$, but does not depend on the network geometry beyond these cost scales.

5.2 Regret Lower Bound

In Theorem 5.3, we establish a tight lower bound matching our upper bound up to logarithmic factors.

Theorem 5.3. *Fix $n \geq 2$ and time horizon $T \geq 2$. Consider the model in Section 3. For any online algorithm* `ALG` *that observes censored demand feedback and $\boldsymbol{P}_t$, there exists an instance (with bounded i.i.d. $(\boldsymbol{d}_t, \boldsymbol{P}_t)$ and linear costs) such that, against the best base-stock policy on Δ_{n-1},*

$$\inf_{\mathit{ALG}} \sup_{\text{instances}} \mathbb{E}\left[\text{Regret}(T)\right] \geq cT^{\frac{n}{n+1}},$$

for a constant $c > 0$ that depends only on $U, \max_{ij} l_{ij}, \max_{ij} c_{ij}$, not on T. The same bound holds for the pseudoregret.

To prove the lower bound, we construct a family of "smoothed discrete" instances with $\boldsymbol{P}_t \equiv I_n$, $l_{ii} \equiv 1$, $c_{ij} \equiv 0$, and i.i.d. demands of the form $d_{t,i} = \eta_{t,i}$ or $d_{t,i} = \theta_i + \eta_{t,i}$ with probability 1/2 each, where $\eta_{t,i}$ are truncated mean-zero Gaussians . Under any static base-stock $\boldsymbol{s} \in \Delta_{n-1}$, the state resets in one step to $\boldsymbol{s}$, so the long-run average modified cost reduces to $\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) = -\sum_{i=1}^{n} \mathbb{E}\min\{d_{t,i}, s_i\} + \text{const}$. A cone-envelope argument shows that $\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta})$ is sandwiched between positive constants times $\|\boldsymbol{s} - \boldsymbol{\theta}\|_1$, up to an $O(n\sigma)$ smoothing error, which we make negligible by choosing the noise level σ sufficiently small. We then take a maximal r-packing $\mathcal{V} \subset \Delta_{n-1}$ with $|\mathcal{V}| \gtrsim r^{-(n-1)}$ and place a uniform prior on $\boldsymbol{\Theta} \in \mathcal{V}$.

Letting $\widehat{\boldsymbol{s}}$ be the empirical average of the actions played by the algorithm, convexity of $\lambda(\cdot \mid \boldsymbol{\theta})$ implies that the (pseudo)regret is at least of order $T\,\mathbb{E}\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1$. A localized KL lemma for censored observations shows that each round carries non-negligible information about $\boldsymbol{\Theta}$ only when the chosen base-stock is within distance $O(r)$ of the true parameter. Combined with the packing geometry, this yields a mutual information bound $I(\boldsymbol{\Theta}; Z_{1:T}) = O(Tr^{n+1})$, while the metric entropy satisfies $\log|\mathcal{V}| \gtrsim (n-1)\log(1/r)$. A metric version of Fano's inequality then implies $\mathbb{E}\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1 \gtrsim r$ unless Tr^{n+1} dominates $\log(1/r)$, so the expected pseudoregret is at least $\Omega(rT)$ in the nontrivial regime.

Optimizing by choosing $r \asymp (\log T/T)^{1/(n+1)}$ yields $\mathbb{E}[\text{PseudoRegret}(T)] \geq c\,T^{\frac{n}{n+1}}$ up to logarithmic factors. The gap between regret and pseudoregret is $O(\sqrt{T \log T})$, which is dominated for $n \geq 2$.

6 NUMERICAL EXPERIMENTS

We consider a simplified setting with uniform repositioning costs $c_{ij} = 1$; in this case the minimum-cost flow solution simplifies to $\|\boldsymbol{y}_t - \boldsymbol{x}_t\|_1/2$, avoiding the need to solve an LP at each step. The lost-sales cost is set uniformly to $l_{ij} = 10$ for all i, j. Demand arrives independently as $d_{t,i} \sim \text{Poisson}(\mu_i)$, with location heterogeneity captured by $\mu_i = \mu_{\min} + \frac{i-1}{n-1}(\mu_{\max} - \mu_{\min})$ for $n > 1$ and $\mu_{\min}$ for $n = 1$; we set $(\mu_{\min}, \mu_{\max}) = (0.2, 0.8)$. Fulfilled items are stochastically redistributed via a row-stochastic routing matrix P_t whose rows are i.i.d. Dirichlet$(1, \ldots, 1)$.

We compare the following three policies: `LipBR`: run over a uniform simplex grid of resolution $m = 200$ with confidence bound scale parameter $H = 5.0$; `NoRepo`: $\boldsymbol{y}_t = \boldsymbol{x}_t$ (no repositioning); and `Uniform`: $\boldsymbol{y}_t = (1/n, \ldots, 1/n)$ (always reposition to a uniform allocation). We performed 20 independent runs for network sizes $n = 2, 3, 4$ and horizons $T = 1000, 2000, 3000$. Table 1 reports the average per-period cost (total cost divided by T) for each policy; lower values indicate better performance. Code can be found at **`https://github.com/hanshengjiang/Lipschitz-NetworkInventory`**.

Table 1: Average per-period cost across network sizes and horizons.

n	Policy	$T = 1000$	$T = 2000$	$T = 3000$
2	`LipBR`	6.322	6.363	6.409
2	`NoRepo`	7.230	7.289	7.306
2	`Uniform`	6.473	6.487	6.506
3	`LipBR`	10.533	10.593	10.537
3	`NoRepo`	12.041	11.976	11.925
3	`Uniform`	11.500	11.439	11.398
4	`LipBR`	15.240	15.413	15.255
4	`NoRepo`	16.787	16.883	16.723
4	`Uniform`	16.355	16.475	16.334

Performance Comparison. Across all n and T, `LipBR` achieves roughly an 11% average reduction in cost versus `NoRepo` and about 5% versus `Uniform`. The gains over `Uniform` become more pronounced as the network size grows ($n \geq 3$), suggesting that the benefits of learning-based repositioning amplify in larger, more heterogeneous systems.

Dependence on T. The reported numbers are average per-period costs averaged over 20 independent runs for each (n, T) pair. Because each horizon is simulated separately and performance is estimated via Monte Carlo, there is no reason to expect these empirical averages to be monotone in T. The small fluctuations (on the order of 1–2%) across horizons are consistent with sampling variability and the slow $T^{-1/(n+1)}$ decay rate predicted by our regret analysis. Importantly, `LipBR` consistently outperforms both `NoRepo` and `Uniform` for all horizons and network sizes.

Discussion. `LipBR` consistently outperforms both baselines across all horizons and network sizes. It dominates `Uniform` without relying on prior knowledge of demand parameters or network symmetry, and the gap relative to `NoRepo` quantifies the value of active repositioning under censored demand. Although these experiments are limited to small networks for rapid validation, our theoretical regret guarantees hold for arbitrary n. These results empirically confirm that `LipBR` correctly exploits the structure of the censored feedback problem and translates theoretical guarantees into tangible cost savings.

7 CONCLUSION

We introduce a Lipschitz bandit perspective into an continuous-state average-cost MDP with censored feedback and correlated transitions. By carefully exploiting the Lipschitz continuity of the long-run cost and developing customized concentration bounds, we designed an algorithm that learns near-optimal repositioning policies with provably sublinear regret. In particular, our regret guarantee $O(T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}})$ not only generalizes the classic $\sqrt{T}$ rate for single-location inventory systems, but also matches a new lower bound that underscores the inherent difficulty of managing an n-location network. These results demonstrate that efficient learning is possible even in complex multi-dimensional MDPs, as long as one can identify and leverage the right structural properties (in our case, Lipschitz continuity) in place of more restrictive assumptions like convexity or independence.

Our work suggests several future directions. One direction is to apply this Lipschitz-based online learning approach to other domains with continuous state or action spaces, such as routing and scheduling problems. Another important question is whether more expressive policy classes beyond base-stock policies can be handled with similar techniques under more complex inventory settings, for example, with lead time. More broadly, this work provides a rigorous framework for developing new, provably efficient algorithms at the intersection of machine learning and operations research.

Acknowledgements

The authors thank the anonymous reviewers and area chair for their helpful comments and suggestions. Hansheng Jiang acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN 2024 05796).

References

Agrawal, S. and Jia, R. (2022). Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Operations Research*, 70(3).

Akturk, D., Candogan, O., and Gupta, V. (2025). Managing resources for shared micromobility: Approximate optimality in large-scale systems. *Management Science*, 71(7):5676–5695.

Benjaafar, S., Jiang, D., Li, X., and Li, X. (2022). Dynamic inventory repositioning in on-demand rental networks. *Management Science*, 68(11):7861–7878.

Gong, X.-Y. and Simchi-Levi, D. (2024). Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science*, 70(9):6139–6157.

He, L., Hu, Z., and Zhang, M. (2020). Robust repositioning for vehicle sharing. *Manufacturing & Service Operations Management*, 22(2):241–256.

Hihat, M., Gaïffas, S., Garrigos, G., and Bussy, S. (2023). Online inventory problems: beyond the iid setting with online convex optimization. *Advances in Neural Information Processing Systems*, 36:20421–20440.

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

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600.

Jiang, H., Sun, C., and Shen, Z.-J. M. (2022). Spatial supply repositioning with censored demand data. *Available at SSRN 4140449*.

Kar, A. and Singh, R. (2025). Provably adaptive average reward reinforcement learning for metric spaces. In *The 41st Conference on Uncertainty in Artificial Intelligence*.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690.

Ortner, R. and Ryabko, D. (2012). Online regret bounds for undiscounted continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 25.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Scarf, H. (1960). The optimality of (s, s) policies in the dynamic inventory problem. *Mathematical Methods in the Social Sciences*.

Sinclair, S. R., Banerjee, S., and Yu, C. L. (2019). Adaptive discretization for episodic reinforcement learning in metric spaces. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–44.

Song, Z. and Sun, W. (2019). Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475*.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Yuan, H., Luo, Q., and Shi, C. (2021). Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science*.

Zipkin, P. (2008). On the structure of lost-sales inventory models. *Operations research*, 56(4):937–944.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Learning in Continuous State-Space MDPs for Network Inventory Management Supplementary Materials

A Proofs of Concentration Inequality

A.1 Proof Organization and Preliminaries

Our main goal of this section is to prove Lemma 4.6, Lemma 4.7, and lastly the main concentration inequality Lemma 4.5 and Corollary 4.8.

Let us walk through the proof roadmap. We first define the finite horizon value function in Definition A.1 and prove that the finite horizon value function has bounded differences starting from different initial states in Lemma A.2. Recalling the definition of loss and bias in Definition 4.1, we use this result to prove that the loss function does not depend on the initial states in Lemma A.3 and the bias function also has bounded difference starting from different initial states in Lemma A.4, respectively.

After that, we use a known relation between the loss and bias derived in Puterman (2014) (restated in our Lemma 4.4) to prove a concentration bound between the finite sum of expected costs and loss in Lemma 4.6. We then prove a concentration bound between the finite sum of observed costs and expected costs in Lemma 4.7. Finally, we combine the above two concentration bounds to prove Lemma 4.5, which is a concentration bound between the finite sum of observed costs and the loss.

Definition A.1 (Finite Horizon Value Function)**.** For any $\boldsymbol{x} \in \Delta_{n-1}$ the finite horizon value function $V_T^{\boldsymbol{S}}(\boldsymbol{x})$ in time T of MDP $\mathcal{M}(\boldsymbol{S}, \boldsymbol{x})$ is defined as:

$$V_T^{\boldsymbol{S}}(\boldsymbol{x}) := \mathbb{E}\left[\sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) \middle| \boldsymbol{x}_1 = \boldsymbol{x}\right].$$

Lemma A.2. *For any $\boldsymbol{S} \in \Delta_{n-1}$ and $T > 0$, and $\boldsymbol{x}, \boldsymbol{x}' \in \Delta_{n-1}$,*

$$V_T^{\boldsymbol{S}}(\boldsymbol{x}) - V_T^{\boldsymbol{S}}(\boldsymbol{x}') \leq 2 \max_{i,j}\{c_{ij}\}.$$

Proof of Lemma A.2. Only the repositioning cost on the first day is different between $V_T^{\boldsymbol{S}}(\boldsymbol{x})$ and $V_T^{\boldsymbol{S}}(\boldsymbol{x}')$ since the lost sales cost $\tilde{L}$ does not depend on $\boldsymbol{x}$ and only the demand and base stock level $\boldsymbol{S}$, therefore

$$V_T^{\boldsymbol{S}}(\boldsymbol{x}) - V_T^{\boldsymbol{S}}(\boldsymbol{x}') = M(\boldsymbol{S} - \boldsymbol{x}) - M(\boldsymbol{S} - \boldsymbol{x}').$$

Let $\mathbf{1}_1 \in \Delta_{n-1}$ denote the inventory level that has all inventory 1 at location 1 and zero elsewhere, then by the sub-additivity of the repositioning cost (see Benjaafar et al. (2022, Lemma 2.3) for a proof), we have

$$\begin{aligned} M(\boldsymbol{S} - \boldsymbol{x}) - M(\boldsymbol{S} - \boldsymbol{x}') &\leq M(\boldsymbol{x}' - \boldsymbol{x}) && \text{(A.1)} \\ &\leq M(\boldsymbol{x}' - \mathbf{1}_1) + M(\mathbf{1}_1 - \boldsymbol{x}) && \text{(A.2)} \\ &\leq 2 \max_{i,j}\{c_{ij}\}, && \text{(A.3)} \end{aligned}$$

where (A.1) and (A.2) use sub-additivity, and the last inequality (A.3) is because the total inventory is bounded by 1 and total number of moved inventory from $\boldsymbol{x}$ or $\boldsymbol{x}'$ to $\mathbf{1}_1$ is at most 1. We thus conclude our proof. □

Next, we use the value difference result in Lemma A.2 to show that the loss $\lambda^{\boldsymbol{S}}(\boldsymbol{x})$ is independent of the starting state $\boldsymbol{x} \in \Delta_{n-1}$.

Lemma A.3. *For any* $\boldsymbol{S}, \boldsymbol{x}, \boldsymbol{x}' \in \Delta_{n-1}$,

$$\lambda^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) = \lambda^{\boldsymbol{S}}(\boldsymbol{x}) =: \lambda^{\boldsymbol{S}}.$$

Proof of Lemma A.3. We note that $\lambda^{\boldsymbol{S}}(\boldsymbol{x}) = \lim_{T\to\infty} \frac{1}{T} V_T^{\boldsymbol{S}}(\boldsymbol{x})$. Therefore, by Lemma A.2 and the assumption that both limits exist (see Remark 4.2 in Section 4), we have

$$\left|\lambda^{\boldsymbol{S}}(\boldsymbol{x}) - \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right| = \left|\lim_{T\to\infty} \frac{1}{T} V_T^{\boldsymbol{S}}(\boldsymbol{x}) - \lim_{T\to\infty} \frac{1}{T} V_T^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right| \le \lim_{T\to\infty} \frac{2\max_{i,j}\{c_{ij}\}}{T} = 0.$$

Hence for any $\boldsymbol{x}$, $\boldsymbol{x}' \in \Delta_{n-1}$, $\lambda^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) = \lambda^{\boldsymbol{S}}(\boldsymbol{x})$. □

Lemma A.4. *For any* $\boldsymbol{S}, \boldsymbol{x}, \boldsymbol{x}' \in \Delta_{n-1}$,

$$\beta^{\boldsymbol{S}}(\boldsymbol{x}') - \beta^{\boldsymbol{S}}(\boldsymbol{x}) \le 2\max_{i,j}\{c_{ij}\}.$$

Proof of Lemma A.4. From Lemma A.3, $\lambda^{\boldsymbol{S}}\left(\boldsymbol{x}\right) = \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) = \lambda^{\boldsymbol{S}}$ for all t. Now by definition of $\beta^{\boldsymbol{S}}(\boldsymbol{x})$ and $\beta^{\boldsymbol{S}}(\boldsymbol{x}')$, we have

$$\beta^{\boldsymbol{S}}(\boldsymbol{x}) = \mathbb{E}\left[\lim_{T\to\infty} \sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \lambda^{\boldsymbol{S}} \middle| \boldsymbol{x}_1 = \boldsymbol{x}\right] = \lim_{T\to\infty} V_T^{\boldsymbol{S}}(\boldsymbol{x}) - T\lambda^{\boldsymbol{S}},$$

and

$$\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) = \mathbb{E}\left[\lim_{T\to\infty} \sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \lambda^{\boldsymbol{S}} \middle| \boldsymbol{x}_1 = \boldsymbol{x}'\right] = \lim_{T\to\infty} V_T^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) - T\lambda^{\boldsymbol{S}}.$$

We note that both of the above limits exist (see Remark 4.2 in Section 4), and hence by Lemma A.2,

$$\begin{aligned}
\beta^{\boldsymbol{S}}(\boldsymbol{x}) - \beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) &= \lim_{T\to\infty}\left(V_T^{\boldsymbol{S}}(\boldsymbol{x}) - T\lambda^{\boldsymbol{S}}\right) - \lim_{T\to\infty}\left(V_T^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) - T\lambda^{\boldsymbol{S}}\right) \\
&= \lim_{T\to\infty} V_T^{\boldsymbol{S}}(\boldsymbol{x}) - V_T^{\boldsymbol{S}}\left(\boldsymbol{x}'\right) \\
&\le 2\max_{i,j}\{c_{ij}\}.
\end{aligned}$$

□

Lemma 4.4 is used to construct a martingale different sequence, and thus the following Azuma-Hoeffding inequality (Wainwright, 2019, Corollary 2.20) can be used to establish a concentration inequality.

Lemma A.5 (Azuma-Hoeffding)**.** *Let* $\{(X_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ *be a martingale difference sequence in the sense that, for all* $k \ge 1$,

$$\mathbb{E}[|X_k|] < \infty \text{ and } \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0. \tag{A.4}$$

Suppose there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ *such that* $X_k \in [a_k, b_k]$ *almost surely for all* $k = 1, \ldots, n$. *Then, for all* $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} X_k\right| \ge \epsilon\right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^{n}(a_k - b_k)^2}\right). \tag{A.5}$$

A.2 Proof of Lemma 4.6 and Proof of Lemma 4.7

Proof of Lemma 4.6. We have the following decomposition.

$$\left|\left(\frac{1}{N}\sum_{t=1}^{N} \mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)\right) - \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)\right| \tag{A.6}$$

$$= \left|\frac{1}{N}\sum_{t=1}^{N}\left(\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)\right)\right| \tag{A.7}$$

$$= \left|\frac{1}{N}\sum_{t=1}^{N} \mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \left(\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) + \mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x}_t,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right] - \beta^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)\right)\right| \tag{A.8}$$

$$=\left|\frac{1}{N}\sum_{t=1}^{N}\left(\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)-\mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x}_t,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right]\right)\right| \tag{A.9}$$

$$=\left|\frac{1}{N}\left(\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)-\mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x}_N,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right]\right)\right. \tag{A.10}$$

$$\left.+\frac{1}{N}\sum_{t=1}^{N-1}\left(\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_{t+1}\right)-\mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x}_t,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right]\right)\right| \tag{A.11}$$

$$\leq\frac{2\max\{c_{ij}\}}{N}+\left|\frac{1}{N}\sum_{t=1}^{N-1}\left(\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_{t+1}\right)-\mathbb{E}_{\boldsymbol{x}_{t+1}\sim\mathbb{P}(\cdot|\boldsymbol{x}_t,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_{t+1}\right)\right]\right)\right| \tag{A.12}$$

Notice that equation (A.7) comes from the uniform loss result of Lemma A.3 that the loss function does not depend on the initial states. By Lemma 4.4 we get equation (A.8). By Lemma A.4 we know that for any $\boldsymbol{S},\boldsymbol{x},\boldsymbol{x}'\in\Delta_{n-1}$, $\beta^{\boldsymbol{S}}(\boldsymbol{x}')-\beta^{\boldsymbol{S}}(\boldsymbol{x})\leq 2\max_{i,j}\{c_{ij}\}$, the difference between any two biases starting from two different states are bounded. Since if we start from a state which is in the state space $\boldsymbol{x}_1\in\Delta_{n-1}$ all the following states are also in the state space Δ_{n-1}, we have that $|\left(\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)-\mathbb{E}_{\boldsymbol{x}'\sim\mathscr{P}^{\boldsymbol{S}}(\boldsymbol{x}_N)}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right]\right)|\leq 2\max_{i,j}\{c_{ij}\}$, therefore we get equation (A.12) by Lemma A.4.

We define the stochastic process $\{\delta_t\}_{t=1}^{N-1}$ as $\delta_{t+1}:=\beta^{\boldsymbol{S}}\left(\boldsymbol{x}_{t+1}\right)-\mathbb{E}_{\boldsymbol{x}'\sim\mathbb{P}(\cdot|\boldsymbol{x}_t,\boldsymbol{S})}\left[\beta^{\boldsymbol{S}}\left(\boldsymbol{x}'\right)\right]$. By the bounded bias result in Lemma A.4, $|\delta_t|\leq 2\max\{c_{ij}\}$ for all t and thus it lies in $[-2\max_{i,j}\{c_{ij}\},-2\max_{i,j}\{c_{ij}\}]$. Additionally, we have $\mathbb{E}\left[\delta_{t+1}|\boldsymbol{x}_t\right]=0$. Therefore, $\{\delta_t\}_{t=2}^{N}$ is a martingale difference sequence with respect to the filtration formed by σ-fields $\sigma(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k)$ for $k=1,\ldots,N$. By Azuma-Hoeffding inequality in Lemma A.5 we have that $\forall\Delta>0$,

$$\mathbb{P}\left(\left|\sum_{t=2}^{N}\delta_t\right|\geq\Delta\right)\leq 2\exp\left(-\frac{\Delta^2}{8(N-1)(\max_{i,j}\{c_{ij}\})^2}\right).$$

Therefore, by setting $\Delta=2\sqrt{2(N-1)\log\left(\frac{4}{\epsilon}\right)}\max_{i,j}\{c_{ij}\}$, we obtain that with probability at least $1-\epsilon/2$,

$$\left|\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)-\lambda^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)\right|\leq\frac{\max_{i,j}\{c_{ij}\}}{N}+\frac{1}{N}2\sqrt{2(N-1)\log\left(\frac{4}{\epsilon}\right)}\max_{i,j}\{c_{ij}\} \tag{A.13}$$

$$\leq\frac{\max_{i,j}\{c_{ij}\}}{N}+2\sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\max_{i,j}\{c_{ij}\} \tag{A.14}$$

□

After constructing the concentration bound of the sum of N expected costs and the loss, we bound the difference between the sum of N expected costs and the sum of N observed costs.

Proof of Lemma 4.7. For $t=1,\ldots,N$, let $\mathcal{F}_t$ denote the σ-field formed by $\{(\boldsymbol{x}_\tau,\boldsymbol{d}_\tau,\boldsymbol{P}_\tau)\}_{\tau=1}^{t}$ and let $\mathcal{F}_0$ be the σ-field formed by $\boldsymbol{x}_1$. By the definition in (6), the cost $C_t^{\boldsymbol{S}}$ is decided by $\boldsymbol{S},\boldsymbol{x}_t,\boldsymbol{d}_t,\boldsymbol{P}_t$ and thus it is measurable with respect to $\mathcal{F}_t$. Furthermore, we define $\delta_t=\widetilde{C}_t^{\boldsymbol{S}}-\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t)$, then δ_t is also measurable with respect to $\mathcal{F}_t$. The expectation $\mathbb{E}[|\delta_t|]\leq\mathbb{E}[|\widetilde{C}_t^{\boldsymbol{S}}|+|\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t)|]\leq 2(2\max_{i,j}\{c_{ij}\}+nU\max_{i,j}l_{ij})=:H$. Clearly, $\mathbb{E}[\delta_1|\mathcal{F}_0]=0$. Furthermore, because $\boldsymbol{x}_t$ is decided by $\boldsymbol{x}_{t-1},\boldsymbol{d}_{t-1},\boldsymbol{P}_{t-1}$ due to the state update equation (1) and the randomness in $C_t^{\boldsymbol{S}}$ comes from $\boldsymbol{x}_t,\boldsymbol{d}_t,\boldsymbol{P}_t$, we have for $t=2,\ldots,N$,

$$\mathbb{E}[\widetilde{C}_t^{\boldsymbol{S}}|\mathcal{F}_{t-1}]=\mathbb{E}[\widetilde{C}_t^{\boldsymbol{S}}|\boldsymbol{x}_{t-1},\boldsymbol{d}_{t-1},\boldsymbol{P}_{t-1}]=\mathbb{E}[\widetilde{C}_t^{\boldsymbol{S}}|\boldsymbol{x}_t]=\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t).$$

Therefore,

$$\mathbb{E}[\delta_t|\mathcal{F}_{t-1}]=\mathbb{E}[\mathbb{E}[\widetilde{C}_t^{\boldsymbol{S}}|\mathcal{F}_{t-1}]-\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t)|\mathcal{F}_{t-1}]=\mathbb{E}[\widetilde{C}^{\boldsymbol{S}}(\boldsymbol{x}_t)-\mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t)|\mathcal{F}_{t-1}]=0.$$

We thus have shown that $\{\delta_t\}_{t=1}^{N}$ is a martingale difference sequence. By the Azuma-Hoeffding inequality in Lemma A.5, we have

$$\mathbb{P}\left(\left|\sum_{t=2}^{N}\delta_t\right|\geq\Delta\right)\leq 2\exp\left(-\frac{\Delta^2}{2(N-1)H^2}\right),$$

where $H = 4\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}$. Therefore, by setting $\Delta = \sqrt{2(N-1)\log\left(\frac{4}{\epsilon}\right)}H$, we obtain that with probability at least $1-\epsilon/2$,

$$\left|\frac{1}{N}\sum_{t=1}^{N}\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \frac{1}{N}\sum_{t=1}^{N}\widetilde{C}_t^{\boldsymbol{S}}\right| \leq \sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}H. \tag{A.15}$$

□

Lemma 4.7 shows the concentration bound of the difference between N observed costs and the expected costs for any given base stock level $\boldsymbol{S}$. Therefore, we can construct a concentration bound for N observed costs and the loss of any policy by combining Lemma 4.6 and Lemma 4.7.

A.3 Proof of Lemma 4.5 and Proof of Corollary 4.8

Proof of Lemma 4.5. According to Lemma 4.6, Lemma 4.7 and the union bound inequality, we have that with probability $1-\epsilon$,

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}C_t^{\boldsymbol{S}} - \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)\right| \leq& \frac{1}{N}\left|\sum_{t=1}^{N}C_t^{\boldsymbol{S}} - \sum_{t=1}^{N}C^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)\right| + \frac{1}{N}\left|\sum_{t=1}^{N}C^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right) - \lambda^{\boldsymbol{S}}\left(\boldsymbol{x}_1\right)\right| \\ \leq& \frac{\max_{i,j}\{c_{ij}\}}{N} + 2\sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\max_{i,j}\{c_{ij}\} \\ &+ \sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\left[4\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}\right] \\ =& \frac{\max_{i,j}\{c_{ij}\}}{N} + \sqrt{\frac{2}{N}\log\left(\frac{4}{\epsilon}\right)}\left[6\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}\right].\end{aligned}$$

□

Proof of Corollary 4.8. Let $B := \frac{\max_{i,j}\{c_{ij}\}}{T} + \sqrt{\frac{2}{T}\log\left(\frac{4}{\epsilon}\right)}\left[6\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}\right]$. On the one hand, we have

$$\min_{\boldsymbol{S}\in\Delta_{n-1}}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}}\,\middle|\,\boldsymbol{x}_1\right] \leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}^*}\,\middle|\,\boldsymbol{x}_1\right] \leq \lambda^{\boldsymbol{S}^*} + B \text{ with probability } 1-\epsilon, \tag{A.16}$$

where the first inequality is by definition of $\min_{\boldsymbol{S}\in\Delta_{n-1}}$ and the second inequality follows from Lemma 4.5.

On the other hand, for some $\boldsymbol{S}^{(T)} \in \Delta_{n-1}$, it holds that

$$\min_{\boldsymbol{S}\in\Delta_{n-1}}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}}\,\middle|\,\boldsymbol{x}_1\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}^{(T)}}\,\middle|\,\boldsymbol{x}_1\right].$$

By Lemma 4.5, it holds with probability $1-\epsilon$ that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}^{(T)}}\,\middle|\,\boldsymbol{x}_1\right] \geq \lambda^{\boldsymbol{S}^{(T)}} - B \geq \lambda^* - B,$$

and thus

$$\min_{\boldsymbol{S}\in\Delta_{n-1}}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}}\,\middle|\,\boldsymbol{x}_1\right] \geq \lambda^* - B. \tag{A.17}$$

Combining (A.16) and (A.17), we have with probability $1-2\epsilon$ that

$$\left|\min_{\boldsymbol{S}\in\Delta_{n-1}}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\widetilde{C}_t^{\boldsymbol{S}}\,\middle|\,\boldsymbol{x}_1\right] - \lambda^*\right| \leq B.$$

Plugging in $\epsilon = T^{-2}/2$ and multiplying both sides of the inequality by a constant T, it follows that with probability $1 - T^{-2}$,

$$\begin{aligned}|\text{Regret}(T) - \text{PseudoRegret}(T)| =& \left| \min_{\boldsymbol{S}\in\Delta_{n-1}} \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{C}_t^{\boldsymbol{S}} \,\middle|\, \boldsymbol{x}_1\right] - T\lambda^* \right| \\ \leq& TB|_{\epsilon=T^{-2}/2} \\ \leq& \max_{i,j}\{c_{ij}\} + \sqrt{2T\log(2) + 4T\log T}\left[6\max_{i,j}\{c_{ij}\} + 2nU\max_{i,j} l_{ij}\right] \\ =& O(\sqrt{T\log T}).\end{aligned}$$

Thus we can conclude our proof. □

B Proofs of Technical Lemmas

B.1 Proof of Lemma 4.9

Proof of Lemma 4.9. We start by bound the covering number of the probability simplex Δ_{n-1} under ℓ_1 norm. Given any $\delta \in (0, \frac{1}{2})$, denote the set of all solutions to (B.1) by $\mathcal{K}_\delta$,

$$k_1 + k_2 + \cdots + k_n = \left\lfloor \frac{1}{\delta} \right\rfloor, k_1, \ldots, k_n \geq 0 \text{ are integers }. \tag{B.1}$$

For any $\boldsymbol{x} \in \Delta_{n-1}$,

$$\left(\left\lfloor \frac{x_1}{\delta} \right\rfloor, \ldots, \left\lfloor \frac{x_{n-1}}{\delta} \right\rfloor, \left\lfloor \frac{1}{\delta} \right\rfloor - \sum_{k=1}^{n-1} \left\lfloor \frac{x_k}{\delta} \right\rfloor \right)$$

is also a solution in $\mathcal{K}_\delta$, and let it be denoted by $(k_1, \ldots, k_n)$.

The ℓ_1 distance between $(x_1, \ldots, x_n)$ and $\delta\boldsymbol{k}$ where $\boldsymbol{k} := (k_1, \ldots, k_n)$ can be bounded as follows,

$$\|\boldsymbol{x} - \delta\boldsymbol{k}\|_1 \leq (n-1)\delta + \delta + (n-1)\delta = (2n-1)\delta.$$

We then bound the cardinality of the set $\mathcal{K}_\delta$. The combinatorial explanation of (B.1) is that: each solution in $\mathcal{K}$ corresponds to a way of inserting $n-1$ dividers to a line of length $\lfloor \frac{1}{\delta} \rfloor + n$. There are $\lfloor \frac{1}{\delta} \rfloor + n - 1$ gaps to put $n-1$ dividers and therefore the total number of distinct nonnegative integer solution to (B.1) is

$$\binom{\lfloor \frac{1}{\delta} \rfloor + n - 1}{n-1}$$

which is at the scale of $O\left(\frac{1}{\delta^{n-1}}\right)$ ignoring multiplicative factors of n that are independent of δ.

Therefore, for $\mathcal{A}_\delta$, it can also be covered by $O\left(\frac{1}{\delta^{n-1}}\right)$ balls of radius δ under ℓ_1 norm.

Lastly, because covering number and packing number are the same up to constant factors (Wainwright, 2019, Lemma 5.5), we conclude that the packing number is also at the scale of $O\left(\frac{1}{\delta^{n-1}}\right)$.

□

B.2 Proof of Lemma 4.10

Proof of Lemma 4.10. We prove the Lipshitz property by comparing the loss of policy S and $S + \boldsymbol{\Delta}$ for any $\boldsymbol{\Delta}$ such that $\sum_i \Delta_i = 0$ and $\|\boldsymbol{\Delta}\|_1 = \delta$. Since the loss doesn't depend on the initial states, we will compare the value of $\lambda^{\boldsymbol{S}}(\boldsymbol{S})$ and $\lambda^{\boldsymbol{S}+\boldsymbol{\Delta}}(\boldsymbol{S} + \boldsymbol{\Delta})$. Recall that the loss is defined in definition 4.1 as follows:

$$\lambda^{\boldsymbol{S}}(\boldsymbol{x}) := \mathbb{E}\left[\lim_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \mathcal{C}^{\boldsymbol{S}}(\boldsymbol{x}_t) \mid \boldsymbol{x}_1 = \boldsymbol{x}\right]$$

where $\boldsymbol{x}$ is the initial state and $\boldsymbol{x}_t$ is the state in the beginning of each time period. Notice that since at the end of each time period, we reposition the system to the same inventory level, the expected cost in each time period

$\mathbb{E}\left[\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)\right]$ is exactly the same as the loss in the first time period. Therefore, we only need to compare the loss in the first time period. Recall that

$$\begin{aligned}\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{x}_t\right)=\mathbb{E}\left[\widetilde{C}_t^{\boldsymbol{S}}(\boldsymbol{x}_t)\;\middle|\;\boldsymbol{x}_t\right]=&\mathbb{E}\left[\widetilde{L}(\boldsymbol{S},\boldsymbol{d}_t,\boldsymbol{P_t})+M(\boldsymbol{S}-\boldsymbol{x}_t)\;\middle|\;\boldsymbol{x}_t\right]\\=&\mathbb{E}\left[L(\boldsymbol{S},\boldsymbol{d}_t,\boldsymbol{P_t})-\sum_{i=1}^n\sum_{j=1}^n l_{ij}P_{t,ij}d_{t,i}+M(\boldsymbol{S}-\boldsymbol{x}_t)\;\middle|\;\boldsymbol{x}_t\right]\end{aligned}\tag{B.2}$$

Since the expectation of the second term $\mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n l_{ij}P_{t,ij}d_{t,i}\right]$ is the same for any policy,

$$\begin{aligned}\mathcal{C}^{\boldsymbol{S}+\boldsymbol{\Delta}}\left(\boldsymbol{S}+\boldsymbol{\Delta}\right)-\mathcal{C}^{\boldsymbol{S}}\left(\boldsymbol{S}\right)=&\,\mathbb{E}\left[L(\boldsymbol{S}+\boldsymbol{\Delta},\boldsymbol{d}_t,\boldsymbol{P}_t)+M(\boldsymbol{S}+\boldsymbol{\Delta}-\boldsymbol{x}_t)\;\middle|\;\boldsymbol{x}_t=[\boldsymbol{S}+\boldsymbol{\Delta}-\boldsymbol{d}_{t-1}]^++\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S},\boldsymbol{d}_t)\right]\\&-\mathbb{E}\left[L(\boldsymbol{S},\boldsymbol{d}_t,\boldsymbol{P}_t)+M(\boldsymbol{S}+\boldsymbol{\Delta}-\boldsymbol{x}_t)\;\middle|\;\boldsymbol{x}_t=[\boldsymbol{S}-\boldsymbol{d}_{t-1}]^++\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S},\boldsymbol{d}_t)\right]\end{aligned}\tag{B.3}$$

For the lost sales cost,

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n l_{ij}\cdot P_{t,ij}(d_{t,i}-S_i)^+-\sum_{i=1}^n\sum_{j=1}^n l_{ij}\cdot P_{t,ij}(d_{t,i}-S_i-\Delta_i)^+\right]\\\le&\sum_{i=1}^n\sum_{j=1}^n l_{ij}P_{t,ij}\Delta_i\\=&\sum_{i=1}^n l_{ij}\Delta_i\le\|\boldsymbol{\Delta}\|_1\max l_{ij},\end{aligned}$$

where the first inequality is because of the Lipschitz property of the function $x^+:=\max\{x,0\}$

For the repositioning costs, the differences are bounded by

$$\begin{aligned}&M\left(\boldsymbol{\Delta}+[\boldsymbol{S}+\boldsymbol{\Delta}-\boldsymbol{d}_{t-1}]^++\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S}+\boldsymbol{\Delta},\boldsymbol{d}_t)-[\boldsymbol{S}-\boldsymbol{d}_{t-1}]^+-\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S},\boldsymbol{d}_t)\right)\\\le&6\|\boldsymbol{\Delta}\|_1\max_{ij}c_{ij},\end{aligned}$$

because one can show that $M(\boldsymbol{x})\le 2\max_{ij}c_{ij}\|\boldsymbol{x}\|_1$ and by Lipschitz properties of $[x]^+$ and min,

$$\|\boldsymbol{\Delta}+[\boldsymbol{S}+\boldsymbol{\Delta}-\boldsymbol{d}_{t-1}]^++\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S}+\boldsymbol{\Delta},\boldsymbol{d}_t)-[\boldsymbol{S}-\boldsymbol{d}_{t-1}]^+-\boldsymbol{P}_t^{\mathrm{T}}\min(\boldsymbol{S},\boldsymbol{d}_t)\|_1\le 3\|\boldsymbol{\Delta}\|_1.$$

To conclude, we have

$$\lambda^{\boldsymbol{S}+\boldsymbol{\Delta}}(\boldsymbol{S}+\boldsymbol{\Delta})-\lambda^{\boldsymbol{S}}(\boldsymbol{S})\le(\max_{ij}l_{ij}+6\max_{ij}c_{ij})\|\boldsymbol{\Delta}\|_1$$

□

C Proof of Regret Upper Bound

Proof of Theorem 5.1. From inequality (11), we can rewrite the pseudoregret as

$$\text{PseudoRegret}=\sum_{t=1}^T\mathbb{E}[\widetilde{C}_t(\boldsymbol{x}_t)]-T\lambda^*\le\underbrace{\sum_{t=1}^T\mathbb{E}[\widetilde{C}'_t(\boldsymbol{x}_t)]-T\lambda^*}_{\text{Regret Part (I)}}+\underbrace{\sum_{t=2}^T\mathbb{1}\{a_{l-1}\neq a_l\}M(\boldsymbol{m}_{a_l}-\boldsymbol{x}_t)}_{\text{Regret Part (II)}}.\tag{C.1}$$

We divide our proof into the following four steps.

Step 1. We establish a few key concentration inequalities based on the technical lemmas proved in Appendix A. For notational simplicity, we define $R(t;k)=\sum_{\tau=1}^t\mathbb{1}\{a_\tau=k\}\mathbb{E}[\widetilde{C}'_\tau(\boldsymbol{x}_\tau)]-t\lambda^*$ be the total regret of all the time periods that arm k is pulled till time t. Let $n_t(k) = \sum_{\tau=1}^{t} \mathbb{1}\{a_\tau = k\}$ be the total number of time periods that arm k is pulled till time t. Let $\Sigma_t(k) = \sum_{\tau=1}^{t} \mathbb{1}\{a_\tau = k\}\tilde{C}'_\tau(\boldsymbol{x}_\tau)$ be the sum of costs of arm k till time t.

According to Lemma 4.5, by setting $\epsilon = T^{-2}$, and

$$H = 2(6 \max_{i,j} c_{ij} + 2nU \max_{i,j} l_{ij}),$$

we have that with probability $1 - T^{-2}$, the sum of observed costs and the loss for arm k,

$$\left| \frac{1}{n_t(k)} \Sigma_t(k) - \lambda^{\boldsymbol{S}_k}(\boldsymbol{x}_1) \right| \leq H \sqrt{\frac{\log(T)}{n_t(k)}}. \tag{C.2}$$

Similarly, this also applies to the best arm k^*, with probability $1 - T^{-2}$ it holds that

$$\left| \frac{1}{n_t(k^*)} \Sigma_t(k^*) - \lambda^{\boldsymbol{S}_{k^*}}(\boldsymbol{x}_1) \right| \leq H \sqrt{\frac{\log(T)}{n_t(k^*)}}. \tag{C.3}$$

Combining inequality (C.2) and (C.3), we have that with probability $1 - 2T^{-2}$

$$-\lambda^{\boldsymbol{S}_k}(\boldsymbol{x}_1) + 2H\sqrt{\frac{\log(T)}{n_t(k)}} \geq -\frac{1}{n_t(k)}\Sigma_t(k) + H\sqrt{\frac{\log(T)}{n_t(k)}}. \tag{C.4}$$

Step 2. We focus on one specific arm k and aim to bound the regret incurred when pulling arm k. We fix t to be the last time that arm k is selected by the upper-confidence-bound rule, i.e., the beginning of the last epoch pulling arm k. Note that the number of times that an arm is used in one epoch doubles every time it is selected, the total number of times arm k is selected before (and including) the last time is always equal to than the epoch length of the last time it is pulled. For example, suppose the algorithm selects arm k four times, then the four epochs' lengths are $1, 2, 4, 8$ respectively. In this case, $n_T(k) = 15$; the last time t arm k is pulled satisfies $n_t(k) = \frac{n_T(k)+1}{2} = 8$. The epoch length of the last time arm k is pulled is also 8. Note that at this time arm k's upper confidence bound is higher than the upper confidence bound of the best arm k^*, therefore we have when t satisfies $n_t(k) = \frac{n_T(k)+1}{2}$,

$$-\frac{1}{n_t(k)}\Sigma_t(k) + H\sqrt{\frac{\log(T)}{n_t(k)}} \geq -\frac{1}{n_t(k^*)}\Sigma_t(k^*) + H\sqrt{\frac{\log(T)}{n_t(k^*)}} \tag{C.5}$$

Combining inequality (C.4), inequality (C.5), and inequality (C.3), we have that with probability at least $1 - 2T^{-2}$, when t satisfies $n_t(k) = \frac{n_T(k)+1}{2}$,

$$-\lambda^{\boldsymbol{S}_k}(\boldsymbol{x}_1) + 2H\sqrt{\frac{\log(T)}{n_t(k)}} \geq -\frac{1}{n_t(k^*)}\Sigma_t(k^*) + H\sqrt{\frac{\log(T)}{n_t(k^*)}} \geq -\lambda^{\boldsymbol{S}_{k^*}}(\boldsymbol{x}_1) \tag{C.6}$$

According to Lemma 4.6 (setting $\epsilon = 2T^{-2}$), with probability $1 - T^{-2}$,

$$\left| \frac{1}{n_T(k)}\mathbb{E}[\Sigma_T(k)] - \lambda^{\boldsymbol{S}}(\boldsymbol{x}_1) \right| \leq 4\max\{c_{ij}\}\sqrt{\frac{\log(T)}{n_T(k)}}, \tag{C.7}$$

which implies

$$-\lambda^{\boldsymbol{S}}(\boldsymbol{x}_1) \leq -\frac{1}{n_T(k)}\mathbb{E}[\Sigma_t(k)] + 4\max\{c_{ij}\}\sqrt{\frac{\log(T)}{n_T(k)}}. \tag{C.8}$$

Combining inequality (C.6) and (C.8), we have that with probability at least $1 - 3T^{-2}$, when t satisfies $n_t(k) = \frac{n_T(k)+1}{2}$,

$$-\frac{1}{n_T(k)}\mathbb{E}[\Sigma_T(k)] + 4\max\{c_{ij}\}\sqrt{\frac{\log(T)}{n_T(k)}} + 2H\sqrt{\frac{\log(T)}{n_t(k)}} \geq -\lambda^{\boldsymbol{S}_{k^*}}(\boldsymbol{x}_1) \tag{C.9}$$

Note that inequality (C.9) also satisfies for all $n_t(k) \le \frac{n_T(k)+1}{2}$ since it will only make the left-hand side larger, then by setting$n_t(k) = \frac{n_T(k)}{2}$ and we have

$$-\frac{1}{n_T(k)}\mathbb{E}[\Sigma_T(k)] + (4\max\{c_{ij}\} + 2\sqrt{2}H)\sqrt{\frac{\log(T)}{n_T(k)}} \ge -\lambda^{\boldsymbol{S}_{k^*}}(\boldsymbol{x}_1) \tag{C.10}$$

Step 3. On the other hand, with probability $3T^{-2}$, inequality (C.10) does not hold. However, since the cumulative regret $\sum_{t=1}^{T}\mathbb{E}[\widetilde{C}_t'(\boldsymbol{x}_t)] - T\lambda^*$ is at most linear in T, multiplying it by $3T^{-2}$ would only result in a $O(\frac{1}{T})$ term which is negligible. Combining the situations where inequality (C.10) holds and fails respectively, we have that

$$\begin{aligned} R(T;k) &= \Sigma_T(k) - n_T(k)\cdot\lambda^{\boldsymbol{S}_{k^*}} && \text{(C.11)} \\ &\le (1-3T^{-2}) \times n_T(k) \times (4\max\{c_{ij}\} + 2\sqrt{2}H)\sqrt{\frac{\log(T)}{n_T(k)}} + 6T^{-2}O(T) \\ &= O\left(\sqrt{n_T(k)\log T}\right). && \text{(C.12)} \end{aligned}$$

Recall that $\mathcal{A}$ denotes the set of all K arms. Then

$$\text{Regret Part (I)} = \sum_{\boldsymbol{S}_k\in\mathcal{A}} R(T;k) \le O(\sqrt{\log T})\sum_{\boldsymbol{S}_k\in\mathcal{A}}\sqrt{n_T(k)}.$$

Since $f(x) = \sqrt{x}$ is a real concave function and $\sum_{\boldsymbol{S}_k\in\mathcal{A}} n_T(k) = T$, by Jensen's inequality we have $\frac{1}{K}\sum_{\boldsymbol{S}_k\in\mathcal{A}}\sqrt{n_T(k)} \le \sqrt{\frac{1}{K}\sum_{\boldsymbol{S}_k\in\mathcal{A}} n_T(k)} = \sqrt{\frac{T}{K}}$.

Therefore, we have

$$\text{Regret Part (I)} \le O(\sqrt{KT\log T}),$$

where K is the total number of arms.

Step 4. According to Lemma 4.9, the number of arms is $O(\frac{1}{\delta^{n-1}})$. Now, if we consider δ not given, according to Lemma 4.10 we will have a discretization error $O(\delta T)$. Taking $\delta = (\log T/T)^{1/(n+1)}$, then the sum of the discretization error and Regret Part (II) is bounded by $C_1(\max_{i,j} c_{ij} + nU\max_{i,j} l_{ij})T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}} = O\left(T^{n/(n+1)}(\log T)^{1/(n+1)}\right)$.

Regret Part (II) is essentially the regret of using the memory point at the beginning of each epoch. Because the epoch length is doubled each time, and therefore there is at most $\log(T/K)$ epochs for each arm and the total number of epochs is at most $K\log(T/K)$ epochs. Also, $M(\boldsymbol{m}_{a_l} - \boldsymbol{x}_t) \le 2\max c_{ij}$ which is a constant. Therefore, the regret of this part is at most $O(K\log(T/K))$. Regret Part (I) is $O(K\log(T/K)) = O(\frac{1}{\delta^{n-1}}\log(\delta^{n-1}T))$, which is negligible comparing to Regret Part (II). To conclude, the total regret is upper bounded by $O\left(T^{n/(n+1)}(\log T)^{1/(n+1)}\right)$.

By Corollary 4.8, the difference of regret and pseudoregret is at most $O(\sqrt{T\log T})$ and therefore

$$\text{Regret}(T) \le \text{PseudoRegret}(T) + O(\sqrt{T\log T}) = O\left(T^{n/(n+1)}(\log T)^{1/(n+1)}\right) \text{ for } n \ge 1.$$

□

D Proof of Regret Lower Bound

We now show that the dependence on T and the network size n in Theorem 5.1 is optimal up to logarithmic factors.

Proof of Theorem 5.3. Fix $n \ge 2$ and $T \ge 2$. We construct a family of hard instances indexed by points in the simplex Δ_{n-1}, and then apply an information-theoretic argument.

Hard instance family. Let $r \in (0, 1/4]$ be a radius to be specified later. By Lemma 4.9, there exists a maximal r-packing $\mathcal{V} \subset \Delta_{n-1}$ under the ℓ_1-metric with cardinality

$$M := |\mathcal{V}| \asymp r^{-(n-1)}.$$

For each parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathcal{V}$ we define an instance of our MDP as follows. Routing is deterministic and identity:

$$\boldsymbol{P}_t \equiv I_n.$$

Lost-sales and repositioning costs are

$$l_{ii} \equiv 1, \qquad l_{ij} \equiv 0 \ (i \neq j), \qquad c_{ij} \equiv 0 \text{ for all } i, j.$$

The demand process $\{d_{t,i}\}$ is i.i.d. over t and i, and bounded by U as required in Section 3. For each coordinate i,

$$d_{t,i} = \begin{cases} \eta_{t,i}, & \text{with probability } 1/2, \\ \theta_i + \eta_{t,i}, & \text{with probability } 1/2, \end{cases}$$

where $\eta_{t,i}$ are i.i.d. draws from a mean-zero Gaussian $\mathcal{N}(0, \sigma^2)$ truncated to $[-U, U]$, with variance parameter $\sigma = \sigma_T > 0$ chosen later. In particular $|d_{t,i}| < U$ almost surely.

This family of instances satisfies all assumptions in Section 3 for every $\boldsymbol{\theta} \in \mathcal{V}$.

Fix $\boldsymbol{\theta} \in \mathcal{V}$ and a (possibly randomized) algorithm ALG. Let $\boldsymbol{y}_t \in \Delta_{n-1}$ be the action (base-stock vector) chosen by ALG at time t.

With $c_{ij} = 0$ and $\boldsymbol{P}_t \equiv I_n$, the repositioning cost vanishes and the modified cost in (7) simplifies to

$$\widetilde{C}_t(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{d}_t, \boldsymbol{P}_t) = -\sum_{i=1}^{n} \min\{d_{t,i}, y_{t,i}\},$$

which does not depend on the current state $\boldsymbol{x}_t$.

Moreover, the state dynamics in (1) simplify to

$$\boldsymbol{x}_{t+1} = (\boldsymbol{y}_t - \boldsymbol{d}_t)^+ + I_n^\top \min(\boldsymbol{y}_t, \boldsymbol{d}_t) = \boldsymbol{y}_t,$$

so the state simply resets in one step to the chosen base-stock vector. In particular, under any policy that plays a fixed base-stock $\boldsymbol{s} \in \Delta_{n-1}$ at all times, the state is constant, $\boldsymbol{x}_t \equiv \boldsymbol{s}$, and the per-period modified cost is i.i.d. with mean

$$\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) := \mathbb{E}\Big[\widetilde{C}_t(\boldsymbol{x}_t, \boldsymbol{s}, \boldsymbol{d}_t, \boldsymbol{P}_t)\Big] = -\sum_{i=1}^{n} \phi_{\theta_i}(s_i), \qquad \phi_{\theta_i}(s) := \mathbb{E} \min\{d_{t,i}, s\}.$$

Thus the long-run average modified cost under base-stock $\boldsymbol{s}$ equals $\lambda(\boldsymbol{s} \mid \boldsymbol{\theta})$.

Let $\lambda^*(\boldsymbol{\theta}) := \inf_{\boldsymbol{s} \in \Delta_{n-1}} \lambda(\boldsymbol{s} \mid \boldsymbol{\theta})$ denote the optimal average modified cost among (static) base-stock policies for instance $\boldsymbol{\theta}$.

The pseudo-regret of ALG on instance $\boldsymbol{\theta}$ (i.e., regret defined in terms of average modified costs) can then be written as

$$\text{PseudoRegret}_{\boldsymbol{\theta}}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \left(\widetilde{C}_t^{\texttt{ALG}} - \widetilde{C}_t^{\boldsymbol{S}^*(\boldsymbol{\theta})}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \lambda(\boldsymbol{y}_t \mid \boldsymbol{\theta})\right] - T\,\lambda^*(\boldsymbol{\theta}),$$

where $\boldsymbol{S}^*(\boldsymbol{\theta})$ is any minimizer of $\lambda(\cdot \mid \boldsymbol{\theta})$.

Because $s \mapsto \min\{d_{t,i}, s\}$ is concave in s for each fixed $d_{t,i}$, each ϕ_{θ_i} is concave and hence $\lambda(\cdot \mid \boldsymbol{\theta})$ is convex on Δ_{n-1}. Therefore, letting

$$\widehat{\boldsymbol{s}} := \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{y}_t \in \Delta_{n-1},$$

Jensen's inequality yields

$$\frac{1}{T} \sum_{t=1}^{T} \lambda(\boldsymbol{y}_t \mid \boldsymbol{\theta}) \ \geq \ \lambda(\widehat{\boldsymbol{s}} \mid \boldsymbol{\theta}), \quad \text{hence} \quad \text{PseudoRegret}_{\boldsymbol{\theta}}(T) \ \geq \ T\Big(\lambda(\widehat{\boldsymbol{s}} \mid \boldsymbol{\theta}) - \lambda^*(\boldsymbol{\theta})\Big). \tag{D.1}$$

Lemma D.1 (Cone envelope under smoothed discrete demands)**.** *There exist absolute constants $a_1, a_2, C > 0$ (independent of T, σ) such that for all $\boldsymbol{\theta}, \boldsymbol{s} \in \Delta_{n-1}$,*

$$a_1 \big\|\boldsymbol{s} - \boldsymbol{\theta}\big\|_1 - Cn\sigma \;\le\; \lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \;\le\; a_2 \big\|\boldsymbol{s} - \boldsymbol{\theta}\big\|_1 + Cn\sigma. \tag{D.2}$$

In particular, whenever $\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 \ge 4Cn\sigma/a_1$, we have

$$\frac{a_1}{2} \big\|\boldsymbol{s} - \boldsymbol{\theta}\big\|_1 \;\le\; \lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \;\le\; 2a_2 \big\|\boldsymbol{s} - \boldsymbol{\theta}\big\|_1. \tag{D.3}$$

Thus the per-round sub-optimality gap for any base-stock $\boldsymbol{s}$ scales linearly with $\|\boldsymbol{s} - \boldsymbol{\theta}\|_1$, up to an additive perturbation of order $n\sigma$.

Let $\boldsymbol{\Theta}$ denote a random parameter drawn uniformly from $\mathcal{V}$. Using Lemma D.1 and (D.1), for each fixed $\boldsymbol{\theta} \in \mathcal{V}$ we have

$$\lambda(\widehat{\boldsymbol{s}} \mid \boldsymbol{\theta}) - \lambda^*(\boldsymbol{\theta}) \;\ge\; \lambda(\widehat{\boldsymbol{s}} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) - \big(\lambda^*(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta})\big) \;\ge\; a_1 \|\widehat{\boldsymbol{s}} - \boldsymbol{\theta}\|_1 - 2Cn\sigma,$$

since any minimizer satisfies $\lambda^*(\boldsymbol{\theta}) \le \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta})$ and applying (D.2) with $\boldsymbol{s} = \boldsymbol{S}^*(\boldsymbol{\theta})$ yields $|\lambda^*(\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta})| \le Cn\sigma$.

Consequently, combining with (D.1),

$$\mathrm{PseudoRegret}_{\boldsymbol{\theta}}(T) \;\ge\; T\Big(a_1 \|\widehat{\boldsymbol{s}} - \boldsymbol{\theta}\|_1 - 2Cn\sigma\Big).$$

Taking expectation over the random draw $\boldsymbol{\Theta}$ (uniform on $\mathcal{V}$), the internal randomness of `ALG`, and the demand process,

$$\mathbb{E}\big[\mathrm{PseudoRegret}(T)\big] \;\ge\; a_1 T\, \mathbb{E}\big[\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1\big] - 2CnT\sigma, \tag{D.4}$$

where $\mathrm{PseudoRegret}(T)$ denotes $\mathrm{PseudoRegret}_{\boldsymbol{\theta}}(T)$ under the uniform prior on $\mathcal{V}$.

We will choose $\sigma = \sigma_T$ such that $nT\sigma_T = o\big(T^{\frac{n}{n+1}}\big)$ (e.g., $\sigma_T = T^{-2}$), so the additive term $2CnT\sigma_T$ will be negligible for our final scaling.

We now show that censored observations Z_t carry information about $\boldsymbol{\theta}$ only when the action $\boldsymbol{y}_t$ is close to $\boldsymbol{\theta}$ in ℓ_1.

Let Z_t denote all observations collected by `ALG` at time t (censored costs and $\boldsymbol{P}_t$), and let H_{t-1} be the history up to time $t-1$. The following lemma quantifies localization of information under our smoothed discrete demand model.

Lemma D.2 (Localized KL bound under smoothed discrete demands)**.** *There exist constants $c_0, c_1, c_2 > 0$ (depending only on U, σ, n) such that for any round t, any (possibly history-dependent) action $\boldsymbol{y}_t \in \Delta_{n-1}$, any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Delta_{n-1}$, any radius $r \in (0,1)$, and any $\gamma \ge 1$,*

$$\mathrm{KL}\Big(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1}) \,\Big\|\, \mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\Big) \;\le\; c_0 \big\|\boldsymbol{\theta} - \boldsymbol{\theta}'\big\|_1^2 \Big(\mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}\|_1 \le \gamma r\} + \mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}'\|_1 \le \gamma r\} + c_1 e^{-c_2 \gamma^2 r^2}\Big). \tag{D.5}$$

Intuitively, the censored observation at time t is informative about $\boldsymbol{\theta}$ only if the chosen base-stock $\boldsymbol{y}_t$ is within distance γr of either $\boldsymbol{\theta}$ or $\boldsymbol{\theta}'$; otherwise, the KL divergence is exponentially small in $\gamma^2 r^2$.

Because $\mathcal{V}$ is an r-packing, there is a constant c_n depending only on n such that for any $\boldsymbol{y} \in \Delta_{n-1}$ and any $\gamma \ge 1$, the ball $\{\boldsymbol{\vartheta} \in \mathcal{V} : \|\boldsymbol{\vartheta} - \boldsymbol{y}\|_1 \le \gamma r\}$ contains at most c_n points. A simple volumetric argument yields the crude bound $c_n \le (1 + 2\gamma)^{n-1}$: the balls of radius $r/2$ centered at $\mathcal{V}$ are disjoint and all those whose centers are within γr of $\boldsymbol{y}$ lie inside a ball of radius $(\gamma + 1/2)r$ around $\boldsymbol{y}$.

Let $\boldsymbol{\Theta}$ be uniform on $\mathcal{V}$ and let $P_{\boldsymbol{\theta}}^{1:T}$ denote the law of the full observation sequence $Z_{1:T}$ under parameter $\boldsymbol{\theta}$ and algorithm `ALG`. We bound the mutual information $I(\boldsymbol{\Theta}; Z_{1:T})$ by averaging pairwise KL divergences over a carefully chosen collection of "neighbor" parameter pairs.

Fix constants $\beta \ge 1$ and $d \in \mathbb{N}$ (depending only on n). For each $\boldsymbol{\theta} \in \mathcal{V}$ choose a neighbor set $\mathcal{N}(\boldsymbol{\theta}) \subset \mathcal{V}$ such that (i) $\boldsymbol{\theta}' \in \mathcal{N}(\boldsymbol{\theta})$ if and only if $\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}')$ (symmetry); (ii) $|\mathcal{N}(\boldsymbol{\theta})| = d$ for all $\boldsymbol{\theta}$ (the resulting graph on $\mathcal{V}$ is d-regular); (iii) for every edge $(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $r \;\le\; \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1 \;\le\; \beta r$..

Such a construction is standard for geometric packings: connect each point to a fixed number of nearest neighbors and, if necessary, rebalance degrees by duplicating a constant number of vertices; this affects M only by a constant factor and does not change the asymptotic scaling.

Define weights

$$W(\boldsymbol{\theta},\boldsymbol{\theta}') := \begin{cases} \frac{1}{Md}, & \boldsymbol{\theta}' \in \mathcal{N}(\boldsymbol{\theta}), \\ 0, & \text{otherwise.} \end{cases}$$

Then $\sum_{\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{\boldsymbol{\theta}} W(\boldsymbol{\theta},\boldsymbol{\theta}') = 1/M$ for all $\boldsymbol{\theta},\boldsymbol{\theta}'$.

By the information-radius inequality and convexity of KL in the second argument (see, e.g., Lemma 2.6 in Tsybakov (2009)),

$$I(\boldsymbol{\Theta}; Z_{1:T}) = \frac{1}{M}\sum_{\boldsymbol{\theta}\in\mathcal{V}} \mathrm{KL}\left(P_{\boldsymbol{\theta}}^{1:T} \middle\| \frac{1}{M}\sum_{\tilde{\boldsymbol{\theta}}\in\mathcal{V}} P_{\tilde{\boldsymbol{\theta}}}^{1:T}\right) \;\le\; \sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}')\mathrm{KL}\left(P_{\boldsymbol{\theta}}^{1:T}\middle\| P_{\boldsymbol{\theta}'}^{1:T}\right). \tag{D.6}$$

By the chain rule for KL under adaptivity,

$$\mathrm{KL}\left(P_{\boldsymbol{\theta}}^{1:T}\middle\| P_{\boldsymbol{\theta}'}^{1:T}\right) = \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{KL}\left(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1})\middle\|\mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\right)\right],$$

where the expectation is under $P_{\boldsymbol{\theta}}^{1:T}$ and H_{t-1} is the history up to time $t-1$.

Applying Lemma D.2 and using that $\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_1 \le \beta r$ for neighbors, we obtain for each t,

$$\begin{aligned}&\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}')\mathbb{E}\left[\mathrm{KL}\left(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1})\middle\|\mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\right)\right] \\ &\qquad \le c_0\beta^2 r^2 \sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}')\mathbb{E}\Big[\mathbb{1}\{\|\boldsymbol{y}_t-\boldsymbol{\theta}\|_1 \le \gamma r\} + \mathbb{1}\{\|\boldsymbol{y}_t-\boldsymbol{\theta}'\|_1 \le \gamma r\}\Big] + c_1\beta^2 r^2 e^{-c_2\gamma^2 r^2}\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}').\end{aligned}$$

Since $\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}') = 1$, the second term is simply $c_1\beta^2 r^2 e^{-c_2\gamma^2 r^2}$.

For the first term, fix $\boldsymbol{y}_t$ and note that

$$\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}')\mathbb{1}\{\|\boldsymbol{y}_t-\boldsymbol{\theta}\|_1 \le \gamma r\} = \sum_{\boldsymbol{\theta}} \mathbb{1}\{\|\boldsymbol{y}_t-\boldsymbol{\theta}\|_1 \le \gamma r\}\sum_{\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}') = \frac{1}{M}\#\Big\{\boldsymbol{\theta}\in\mathcal{V} : \|\boldsymbol{y}_t-\boldsymbol{\theta}\|_1 \le \gamma r\Big\} \le \frac{c_n}{M},$$

and the same bound holds for the indicator involving $\boldsymbol{\theta}'$. Hence

$$\sum_{\boldsymbol{\theta},\boldsymbol{\theta}'} W(\boldsymbol{\theta},\boldsymbol{\theta}')\mathbb{E}\left[\mathrm{KL}\left(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1})\middle\|\mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\right)\right] \;\le\; \frac{2c_0\beta^2 c_n}{M} r^2 + c_1\beta^2 r^2 e^{-c_2\gamma^2 r^2}.$$

Summing over $t = 1,\dots,T$ and using (D.6) yields

$$I(\boldsymbol{\Theta}; Z_{1:T}) \;\le\; T\left(\frac{2c_0\beta^2 c_n}{M} r^2\right) + T\left(c_1\beta^2 r^2 e^{-c_2\gamma^2 r^2}\right). \tag{D.7}$$

Because the constants c_0, c_1, c_2 depend on σ and we are free to choose $\sigma = \sigma_T$ in our construction, we can make the second term in (D.7) negligible compared to the first for the particular choice of $r = r_T$ made below. Concretely, we choose σ_T sufficiently small so that

$$Tc_1\beta^2 r_T^2 e^{-c_2(\sigma_T)\gamma^2 r_T^2} \le 1 \quad \text{for all large } T,$$

and absorb this additive constant into lower-order terms. For the regret scaling, the dominant contribution to $I(\boldsymbol{\Theta}; Z_{1:T})$ is therefore

$$I(\boldsymbol{\Theta}; Z_{1:T}) \;\lesssim\; \frac{2c_0\beta^2 c_n T}{M} r^2,$$

where $M \asymp r^{-(n-1)}$, so the leading term behaves as Tr^{n+1} up to multiplicative constants and polylogarithmic factors coming from c_n.

We now convert the mutual information bound into a lower bound on the estimation error $\mathbb{E}\big[\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1\big]$ via a metric version of Fano's inequality (see, e.g., Theorem 2.7 in Tsybakov (2009)). Since $\mathcal{V}$ is an r-packing under $\|\cdot\|_1$, for any estimator $\widehat{\boldsymbol{s}}$ taking values in Δ_{n-1},

$$\mathbb{E}\big[\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1\big] \;\geq\; \frac{r}{2}\left(1 - \frac{I(\boldsymbol{\Theta}; Z_{1:T}) + \log 2}{\log M}\right). \tag{D.8}$$

(Indeed, define $\widehat{\boldsymbol{\Theta}}$ as the nearest neighbor of $\widehat{\boldsymbol{s}}$ in $\mathcal{V}$; then $\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1 \geq r/2$ whenever $\widehat{\boldsymbol{\Theta}} \neq \boldsymbol{\Theta}$, and Fano's inequality controls $\mathbb{P}(\widehat{\boldsymbol{\Theta}} \neq \boldsymbol{\Theta})$ in terms of $I(\boldsymbol{\Theta}; Z_{1:T})$ and $\log M$.)

Combining (D.7) and (D.8), using $M \asymp r^{-(n-1)}$ and thus $\log M \asymp (n-1)\log(1/r)$, we obtain

$$\mathbb{E}\big[\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1\big] \;\gtrsim\; \frac{r}{2}\left(1 - \frac{C_1 T r^{n+1} + C_2 + \log 2}{(n-1)\log(1/r)}\right),$$

for suitable constants $C_1, C_2 > 0$ (depending on $n, U, \beta, c_0, c_1, c_n$ but not on T or r). Inserting this into (D.4),

$$\mathbb{E}\big[\mathrm{PseudoRegret}(T)\big] \;\gtrsim\; a_1 T r\left(1 - \frac{C_1 T r^{n+1} + C_2 + \log 2}{(n-1)\log(1/r)}\right) - 2CnT\sigma_T.$$

We now choose $r = r_T$ to balance Tr^{n+1} against $\log(1/r)$. Take

$$r_T \asymp \left(\frac{\log T}{T}\right)^{\frac{1}{n+1}}.$$

Then $Tr_T^{n+1} \asymp \log T$ and $\log(1/r_T) \asymp \log T$, so for T large enough the fraction $(C_1 T r_T^{n+1} + C_2 + \log 2)/\log(1/r_T)$ is bounded strictly below 1, say by 1/2. As a result,

$$\mathbb{E}\big[\|\widehat{\boldsymbol{s}} - \boldsymbol{\Theta}\|_1\big] \;\gtrsim\; r_T \quad \text{and} \quad \mathbb{E}\big[\mathrm{PseudoRegret}(T)\big] \;\gtrsim\; Tr_T - 2CnT\sigma_T.$$

Recall that we chose σ_T so that $nT\sigma_T = o\big(T^{\frac{n}{n+1}}\big)$ (for example, $\sigma_T = T^{-2}$ suffices). Therefore the additive term $2CnT\sigma_T$ is negligible compared to $Tr_T \asymp T^{\frac{n}{n+1}}(\log T)^{\frac{1}{n+1}}$. In particular, for all sufficiently large T there exists a constant $c' > 0$ such that

$$\mathbb{E}\big[\mathrm{PseudoRegret}(T)\big] \;\geq\; c'\, T^{\frac{n}{n+1}}.$$

By Yao's minimax principle, this implies that for every algorithm ALG there exists at least one instance in our family (i.e., at least one $\boldsymbol{\theta} \in \mathcal{V}$) such that the pseudo-regret on that instance is at least $c' T^{\frac{n}{n+1}}$.

Finally, Corollary 4.8 implies that for every $\boldsymbol{\theta}$ and every T,

$$\big|\mathrm{Regret}(T) - \mathrm{PseudoRegret}(T)\big| \;\leq\; \widetilde{O}(\sqrt{T}),$$

where the $\widetilde{O}(\cdot)$ hides polylogarithmic factors. Since $T^{\frac{n}{n+1}} \gg \sqrt{T}$ for all $n \geq 2$, this lower-order term does not affect the scaling. Hence there exists a constant $c > 0$ such that for every T and every algorithm ALG, some instance in our family satisfies

$$\mathbb{E}[\mathrm{Regret}(T)] \;\geq\; c\, T^{\frac{n}{n+1}},$$

which completes the proof of Theorem 5.3. □

Below, we provide proofs of two auxiliary facts, Lemmas D.1 and D.2, used in the proof of Theorem 5.3.

Proof of Lemma D.1. We proceed in two steps. First we analyze a purely discrete benchmark (no smoothing noise), for which the cone shape is exact. Then we show that adding the small truncated Gaussian noise perturbs the cost by at most $O(\sigma)$ per coordinate.

Step 1: Discrete benchmark (no noise). Fix a single coordinate and suppress the index i. Consider the discrete demand

$$d^{\text{disc}} = \begin{cases} 0, & \text{with probability } 1/2, \\ \theta, & \text{with probability } 1/2, \end{cases}$$

and define

$$\phi_\theta^{\text{disc}}(s) := \mathbb{E}\min\{d^{\text{disc}}, s\}, \qquad \lambda^{\text{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) := -\sum_{i=1}^n \phi_{\theta_i}^{\text{disc}}(s_i).$$

Since $s \geq 0$, we have

$$\phi_\theta^{\text{disc}}(s) = \frac{1}{2}\min\{0, s\} + \frac{1}{2}\min\{\theta, s\} = \frac{1}{2}\min\{\theta, s\}.$$

Hence for any s,

$$\phi_\theta^{\text{disc}}(\theta) - \phi_\theta^{\text{disc}}(s) = \frac{1}{2}\big(\theta - \min\{\theta, s\}\big) = \begin{cases} \frac{1}{2}(\theta - s), & s \leq \theta, \\ 0, & s \geq \theta. \end{cases}$$

Now return to the vector case $\boldsymbol{s}, \boldsymbol{\theta} \in \Delta_{n-1}$. Let $\Delta_i := s_i - \theta_i$. Then

$$\begin{aligned} \lambda^{\text{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\text{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}) &= -\sum_{i=1}^n \big(\phi_{\theta_i}^{\text{disc}}(s_i) - \phi_{\theta_i}^{\text{disc}}(\theta_i)\big) = \sum_{i=1}^n \big(\phi_{\theta_i}^{\text{disc}}(\theta_i) - \phi_{\theta_i}^{\text{disc}}(s_i)\big) \\ &= \frac{1}{2}\sum_{i: s_i \leq \theta_i} (\theta_i - s_i) = \frac{1}{2}\sum_{i:\Delta_i \leq 0} |\Delta_i|. \end{aligned}$$

Because $\boldsymbol{s}, \boldsymbol{\theta} \in \Delta_{n-1}$, we have $\sum_{i=1}^n \Delta_i = 0$, so

$$\sum_{i:\Delta_i > 0} \Delta_i = \sum_{i:\Delta_i \leq 0} (-\Delta_i) = \sum_{i:\Delta_i \leq 0} |\Delta_i|.$$

The ℓ_1-distance is

$$\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 = \sum_{i=1}^n |\Delta_i| = \sum_{i:\Delta_i > 0} \Delta_i + \sum_{i:\Delta_i \leq 0} |\Delta_i| = 2\sum_{i:\Delta_i \leq 0} |\Delta_i|.$$

Therefore,

$$\lambda^{\text{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\text{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}) = \frac{1}{4}\big\|\boldsymbol{s} - \boldsymbol{\theta}\big\|_1 \qquad \text{for all } \boldsymbol{s}, \boldsymbol{\theta} \in \Delta_{n-1}. \tag{D.9}$$

Thus the discrete benchmark has an exact cone envelope with slope $1/4$ in $\|\cdot\|_1$.

Step 2: Perturbation by truncated Gaussian noise. We now consider the smoothed discrete Gaussian instance in the theorem statement. For a single coordinate, recall that

$$d = \begin{cases} \eta, & \text{with probability } 1/2, \\ \theta + \eta, & \text{with probability } 1/2, \end{cases} \qquad \eta \sim \mathcal{N}(0, \sigma^2) \text{ truncated to } [-U, U].$$

Define

$$\phi_\theta(s) := \mathbb{E}\min\{d, s\}.$$

We compare ϕ_θ to the discrete benchmark $\phi_\theta^{\text{disc}}$. Using the law of d,

$$\phi_\theta(s) = \frac{1}{2}\mathbb{E}\min\{\eta, s\} + \frac{1}{2}\mathbb{E}\min\{\theta + \eta, s\}.$$

Introduce

$$\phi_\theta^{\text{disc}}(s) = \frac{1}{2}\min\{0, s\} + \frac{1}{2}\min\{\theta, s\} = \frac{1}{2}\min\{\theta, s\},$$

as above. Then

$$\big|\phi_\theta(s) - \phi_\theta^{\text{disc}}(s)\big| \leq \frac{1}{2}\big|\mathbb{E}\min\{\eta, s\} - \min\{0, s\}\big| + \frac{1}{2}\big|\mathbb{E}\min\{\theta + \eta, s\} - \min\{\theta, s\}\big|.$$

For both terms we use that $x \mapsto \min\{x, s\}$ is 1-Lipschitz. Indeed,

$$\big|\min\{\eta, s\} - \min\{0, s\}\big| \le |\eta - 0| = |\eta|, \qquad \big|\min\{\theta + \eta, s\} - \min\{\theta, s\}\big| \le |\theta + \eta - \theta| = |\eta|.$$

Therefore,

$$\big|\phi_\theta(s) - \phi_\theta^{\mathrm{disc}}(s)\big| \le \frac{1}{2}\mathbb{E}|\eta| + \frac{1}{2}\mathbb{E}|\eta| = \mathbb{E}|\eta|.$$

Since η is a truncated $\mathcal{N}(0, \sigma^2)$, we have $\mathbb{E}[\eta^2] \le \sigma^2$, hence by Cauchy–Schwarz

$$\mathbb{E}|\eta| \le \sqrt{\mathbb{E}[\eta^2]} \le \sigma.$$

Thus, for all θ and s,

$$\big|\phi_\theta(s) - \phi_\theta^{\mathrm{disc}}(s)\big| \le \sigma. \tag{D.10}$$

Now return to the vector case. We have

$$\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) = -\sum_{i=1}^{n} \phi_{\theta_i}(s_i), \qquad \lambda^{\mathrm{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) = -\sum_{i=1}^{n} \phi_{\theta_i}^{\mathrm{disc}}(s_i).$$

By (D.10),

$$\big|\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\mathrm{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta})\big| \le \sum_{i=1}^{n} \big|\phi_{\theta_i}(s_i) - \phi_{\theta_i}^{\mathrm{disc}}(s_i)\big| \le n\sigma.$$

The same bound holds with $\boldsymbol{s}$ replaced by $\boldsymbol{\theta}$:

$$\big|\lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) - \lambda^{\mathrm{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta})\big| \le n\sigma.$$

Therefore,

$$\begin{aligned}
\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) &= \big(\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\mathrm{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta})\big) + \big(\lambda^{\mathrm{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\mathrm{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta})\big) + \big(\lambda^{\mathrm{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta})\big) \\
&\ge \lambda^{\mathrm{disc}}(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda^{\mathrm{disc}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}) - 2n\sigma \\
&= \frac{1}{4}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 - 2n\sigma,
\end{aligned}$$

where we used (D.9). Similarly,

$$\lambda(\boldsymbol{s} \mid \boldsymbol{\theta}) - \lambda(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \le \frac{1}{4}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 + 2n\sigma.$$

Thus (D.2) holds with $a_1 = a_2 = \frac{1}{4}$ and $C = 2$. Finally, if $\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 \ge 4Cn\sigma/a_1 = 32n\sigma$, then

$$\frac{1}{4}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 - 2n\sigma \;\ge\; \frac{1}{8}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1,$$

and

$$\frac{1}{4}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 + 2n\sigma \;\le\; \frac{3}{8}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1 \le 2 \cdot \frac{1}{4}\|\boldsymbol{s} - \boldsymbol{\theta}\|_1.$$

Absorbing constants into a_1, a_2 yields the effective cone bounds (D.3). □

Proof of Lemma D.2. We first work in one dimension and then lift the result to the vector case.

Fix a round t and suppress the coordinate index. Conditional on the parameter θ and on the chosen action $y := y_t \in \mathbb{R}$, the demand d is drawn as

$$d = \begin{cases} \varepsilon, & \text{with prob. } 1/2, \\ \theta + \varepsilon, & \text{with prob. } 1/2, \end{cases} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ truncated to } [-U, U].$$

We observe the censored quantity

$$Z := \min\{d, y\}.$$

Let $Q^y_\alpha := \mathcal{L}\big(\min\{\alpha+\varepsilon, y\}\big)$ denote the law of the censored observation when the mean is α (so the "baseline" case is Q^y_0 and the shifted case is Q^y_θ). The density of Q^y_α is a mixture of a truncated Gaussian density on $(-\infty, y)$ and an atom at y; a direct computation using the explicit expressions for these densities and standard Gaussian tail bounds (see, e.g., Appendix A in Tsybakov (2009)) shows that there exist constants $\kappa \in (0,1)$ and $C_0, C_1, C_2 > 0$ such that, for any $r \in (0,1)$ and any $\gamma \geq 1$,

$$\mathrm{KL}\big(Q^y_\theta \,\|\, Q^y_{\theta'}\big) \;\leq\; C_0(\theta-\theta')^2\Big(\mathbb{1}\big\{|y-\theta| \leq \kappa\gamma r\big\} + \mathbb{1}\big\{|y-\theta'| \leq \kappa\gamma r\big\} + C_1 e^{-C_2\gamma^2 r^2}\Big). \tag{D.11}$$

The first two terms correspond to the "local" regime where censoring occurs in a region where the two underlying truncated Gaussians differ significantly; the last term accounts for the exponentially small probability that the Gaussian noise bridges a gap of order γr.

In our smoothed discrete instance, the law of Z given (θ, y) is not Q^y_θ but a mixture of the form

$$P^y_\theta := \mathcal{L}(Z \mid \theta, y) = \tfrac{1}{2}Q^y_0 + \tfrac{1}{2}Q^y_\theta,$$

and similarly $P^y_{\theta'} = \tfrac{1}{2}Q^y_0 + \tfrac{1}{2}Q^y_{\theta'}$.

We now use a simple log-sum inequality: if R, P, Q are probability measures such that P, Q are absolutely continuous with respect to R and we define mixtures

$$\widetilde{P} = (1-\lambda)R + \lambda P, \qquad \widetilde{Q} = (1-\lambda)R + \lambda Q,$$

for some $\lambda \in [0,1]$, then

$$\mathrm{KL}(\widetilde{P}\|\widetilde{Q}) \leq \lambda\, \mathrm{KL}(P\|Q). \tag{D.12}$$

To verify (D.12), let $p, q, r, \tilde{p}, \tilde{q}$ denote the corresponding densities with respect to a common dominating measure. Then

$$\tilde{p} = (1-\lambda)r + \lambda p, \qquad \tilde{q} = (1-\lambda)r + \lambda q.$$

By the log-sum inequality,

$$\tilde{p}\log\frac{\tilde{p}}{\tilde{q}} \leq (1-\lambda)r\log\frac{r}{r} + \lambda p \log\frac{\lambda p}{\lambda q} = \lambda p \log\frac{p}{q},$$

and integrating both sides yields (D.12).

Applying (D.12) with $R = Q^y_0$, $P = Q^y_\theta$, $Q = Q^y_{\theta'}$, and $\lambda = 1/2$ gives

$$\mathrm{KL}\big(P^y_\theta \,\|\, P^y_{\theta'}\big) \leq \frac{1}{2}\,\mathrm{KL}\big(Q^y_\theta \,\|\, Q^y_{\theta'}\big). \tag{D.13}$$

Combining (D.11) and (D.13), and enlarging κ if necessary so that $\{|y-\theta| \leq \kappa\gamma r\}$ implies $\{|y-\theta| \leq \gamma r\}$, we obtain the one-dimensional localized KL bound for the smoothed discrete instance:

$$\mathrm{KL}\big(P^y_\theta \,\|\, P^y_{\theta'}\big) \;\leq\; C_0'(\theta-\theta')^2\Big(\mathbb{1}\big\{|y-\theta| \leq \gamma r\big\} + \mathbb{1}\big\{|y-\theta'| \leq \gamma r\big\} + C_1 e^{-C_2\gamma^2 r^2}\Big), \tag{D.14}$$

for new constants $C_0', C_1, C_2 > 0$.

We now return to the full vector $\boldsymbol{Z}_t$ at round t. Conditional on $\boldsymbol{y}_t$ and H_{t-1}, the coordinates $\{Z_{t,i}\}_{i=1}^n$ are independent, and $\boldsymbol{P}_t$ is independent of $\boldsymbol{\theta}$ (thus contributing zero to the KL divergence). Therefore, by independence and the chain rule for KL,

$$\begin{aligned}&\mathrm{KL}\Big(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1}) \,\Big\|\, \mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\Big)\\ &\qquad = \sum_{i=1}^n \mathbb{E}\Big[\mathrm{KL}\Big(\mathcal{L}(Z_{t,i} \mid \theta_i, y_{t,i}) \,\big\|\, \mathcal{L}(Z_{t,i} \mid \theta_i', y_{t,i})\Big) \;\Big|\; \mathsf{H}_{t-1}\Big].\end{aligned}$$

Applying (D.14) coordinate-wise and using the fact that

$$\{\|\boldsymbol{y}_t - \boldsymbol{\theta}\|_1 \leq \gamma r\} \;\subseteq\; \{|y_{t,i} - \theta_i| \leq \gamma r\} \quad \text{for all } i,$$

we can replace the coordinate-wise indicators by the global ones, obtaining

$$\begin{aligned}\mathrm{KL}\Big(\mathcal{L}(Z_t \mid \boldsymbol{\theta}, \mathsf{H}_{t-1}) \Big\| \mathcal{L}(Z_t \mid \boldsymbol{\theta}', \mathsf{H}_{t-1})\Big) &\le C_0' \sum_{i=1}^{n} (\theta_i - \theta_i')^2 \Big(\mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}\|_1 \le \gamma r\} + \mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}'\|_1 \le \gamma r\} + C_1 e^{-C_2\gamma^2 r^2}\Big) \\ &\le C_0' \big\|\boldsymbol{\theta} - \boldsymbol{\theta}'\big\|_1^2 \Big(\mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}\|_1 \le \gamma r\} + \mathbb{1}\{\|\boldsymbol{y}_t - \boldsymbol{\theta}'\|_1 \le \gamma r\} + C_1 e^{-C_2\gamma^2 r^2}\Big),\end{aligned}$$

since $\sum_{i=1}^{n} (\theta_i - \theta_i')^2 \le \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1^2$. Renaming constants as $c_0 := C_0'$, $c_1 := C_1$, and $c_2 := C_2$ yields the desired bound (D.5). □