Les Cahiers du GERAD ISSN: 0711–2440

Online facility location: Running stores on wheels with spatial demand learning

J. Cao, W. Qi, Y. Zhang

G–2025–45

July 2025

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2025
– Bibliothèque et Archives Canada, 2025

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2025
– Library and Archives Canada, 2025

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Online facility location: Running stores on wheels with spatial demand learning

Junyu Cao [a]

Wei Qi [b, c, d]

Yan Zhang [c, d]

[a] *McCombs School of Business, The University of Texas at Austin, Austin (Tx), United States, 78712*

[b] *Department of Industrial Engineering, Tsinghua University, Beijing, China, 100084*

[c] *Desautels Faculty of Management, McGill University, Montréal (Qc), Canada, H3A 1G5*

[d] *GERAD, Montréal (Qc), Canada, H3T 1J4*

junyu.cao@mccombs.utexas.edu
qiw@tsinghua.edu.cn
yan.zhang13@mail.mcgill.ca

July 2025
Les Cahiers du GERAD
G–2025–45

Abstract : A shift toward shopping at (autonomous) wheeled vending stores is redefining urban retail. Compared with traditional brick-and-mortar stores, such mobile stores are cost-efficient to deploy and adaptive to fast-evolving business environments. However, mobile stores are confronted with unknown demand and limited capacity. Store mobility enables demand learning and profit maximization, yet an optimal dynamic store location policy remains unclear. We model this "learning-and-earning" problem by taking optimistic actions under parameter uncertainty. The joint optimization over parameter and action set is complicated by the combinatorial nature and infinite choices within the action set. We overcome these challenges by leveraging continuous approximation methods, and then propose a continuous-approximation optimistic (CA-O) learning framework under some special problem structures. Nevertheless, for more general scenarios, the problem remains intricate due to the nonconvexity in unknown parameters. We alternatively propose a CA-O faster learning algorithm by utilizing first-order approximation techniques and further proving a closed-form gradient to guarantee computational efficiency. We theoretically analyze and numerically validate the regret for the proposed algorithms. In a Toronto case study, our algorithm significantly outperforms baselines. Mobile stores earn higher profits than brick-and-mortar stores through demand learning and store mobility. More broadly, this paper envisions the future landscape of urban retail enhanced by omnipresent mobile facilities.

Keywords: Mobile retail, facility location, contextual bandits, continuous approximation, joint learning and optimization

1 Introduction

The retail landscape is witnessing a surge of innovation in both in-store and online shopping, brought by autonomous technologies. Unmanned stores are redefining the in-store shopping experience by providing cashierless and automated service to customers (e.g., Amazon Go stores in the US and UK). The application of robotics and self-driving technology in fulfillment and delivery holds great promise for advancing online shopping. Kroger collaborates with Nuro to introduce driverless cars to speed up the adoption of grocery delivery, and Domino's Pizza Inc. and Yum Brands Inc.'s Pizza Hut also are exploring driverless vehicles for pizza deliveries (WSJ 2018).

Figure 1: Sample Mobile Retail Stores. (a) Robomart (2023). (b) Nuro (2023). (c) Neolix (2023).

Autonomous technologies now are spurring the retail industry to evolve further beyond unmanned stores or deliveries. A business model of selling products through automated *stores on wheels* is emerging. In such a business model, the retailer is able to place mobile stores at various locations to leverage demand dynamics and increase profits. For example, Unilever partnered with a startup "Robomart" to deploy a fleet of robotic vehicles to sell ice cream through parts of Los Angeles (Forbes 2022). Robomart launched a flexible platform for retailers to sell goods with running stores (Figure 1(a)). Consumers just walk to a nearby van stocked with merchandise, open the van with a swipe on a phone using their app, and complete purchases via their mobile device. The potential market for mobile retail stores is substantial, and the advances in self-driving technology further stimulate the market. Investors and operators are already investing heavily into self-driving vans, e.g., SoftBank invested $940 million in start-up Nuro in 2019 for driverless retail (FT 2019), whose prototype is shown in Figure 1(b). Not only are companies in the US investigating the business model of mobile retail stores, but overseas companies are also joining the trend. For instance, Neolix, a Beijing-based startup, received the approval to operate their autonomous vehicles in both Europe and Asian countries (Bloomberg 2021). Neolix has successfully deployed their vehicles in various application scenarios, including mobile retail stores, as shown in Figure 1(c).

Mobile retail stores are gaining increasing attention and practice in the industry, but research on their operations remains scarce. The key to success is still a mystery. A thorough analysis of the pros and cons is necessary, particularly since the market for mobile retail stores in cities is still in its infancy. The potential of this new retail channel stems from the following two advantages.

Mobility. Mobile retail stores enable retailers to move their stores as swiftly as relocating a car, in contrast to the stationary nature of brick-and-mortar stores. This adjustability of store locations benefits both retailers and customers. Retailers can increase profits by relocating stores to high-demand regions in a city. For customers, the mobile stores provide an engaging touch-and-feel shopping experience and extra proximity as store locations change.

Cost efficiency. Three factors contribute to the cost efficiency of mobile retail stores. First, mobile stores provide an opportunity to reduce labor costs, as demonstrated by the three robotics-enabled practices in Figure 1. Second, retailers can avoid the heavy investment required for physical real estate. Furthermore, mobile stores offer retailers the freedom to explore and test new markets without having to commit to a permanent location.

Despite being a novel retail channel, mobile retail stores come with their own set of unique obstacles. We consulted experts from Kroger to better understand and address these issues. Two main obstacles, if left unaddressed by the operators, could pose risks to this business.

Unknown demand. The extent of customer demand for the novel retail channel in a city is unknown. If demand is misestimated, retailers will run the risk of reduced profitability if they deploy too many stores in low-demand areas or too few in high-demand areas. The observed demand is subject to noise due to the random nature of daily customer demand. Furthermore, demand fluctuates over time as contextual covariates (such as weather, population density, and fuel price) vary.

Limited capacity. The inventory capacity of mobile retail stores is more restricted than that of brick-and-mortar stores. Increased frequency of inventory replenishment could result in higher replenishment costs in the mobile retail channel. Thus, it is important to pay close attention to the replenishment process and the related costs in supply chains for mobile stores.

Fortunately, these challenges can be addressed by utilizing the advantages mentioned. Retailers are able to place mobile stores at various locations and identify local customer demand via daily sales. Store mobility enables cost-efficient location adjustments to learn potential demand. Retailers further mitigate the effects of limited capacity by developing a data-driven policy to optimize mobile store operations.

Motivated by these operational challenges and opportunities of retail on wheels, this paper examines the sequential location decisions made by a retailer managing a fleet of mobile stores in an online setting. The retailer faces uncertainty in spatial demand distribution and determines store locations based on current demand information. After observing the daily sales of each store, the retailer updates their knowledge of demand and decides on the locations for the following day. In this learning-and-earning environment, the retailer must carefully simultaneously explore (i.e., estimate parameters) and exploit (i.e., maximize profit) over time. Meanwhile, spatial demand learning conditioned on contextual covariates introduces greater complexity because the observation is the aggregate demand expressed at these decided store locations. It is insufficient to simply apply established methodologies. This paper proposes a novel online learning framework to help the retailer find a set of store locations adaptively over the planning horizon.

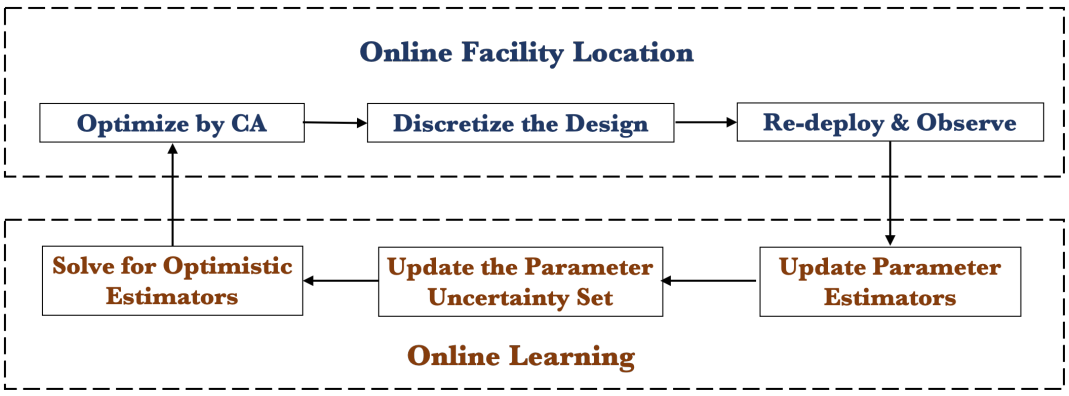


Figure 2: The CA-O Learning Algorithmic Framework.

The main contributions of this paper are summarized as follows:

I. *Formulation*: To the best of our knowledge, this paper is an early attempt to consider a mobile retail store location problem in an online setting with unknown parameters and contextual covariates. We formulate a sequential location problem to maximize the overall profit and analyze the profitability of mobile retail.

II. *Theoretical Contributions:* The major challenges of mobile retail store operations stem from 1) the complexity of action space (i.e., store locations) and 2) the interdependence between actions and observations (i.e., demand). Our framework, visualized in Figure 2, addresses both challenges by simplifying the problem and translating the recipe into discrete location decisions. More specifically:

   a) To balance between exploration and exploitation, we formulate an optimistic optimization problem, in which the retailer selects the plausibly best estimator from a properly constructed uncertainty set for unknown parameters and decides on store locations. The optimistic decision facilitates both the acquisition of information to overcome uncertainty and the maximization of profit.

   b) Selecting store locations is an infinite-dimensional decision problem, much more complicated than a problem of choosing a variable/vector in traditional bandits problems. Faced with an action set of high complexity, we leverage the analytical convenience inherent in facility location models and use continuous approximation (CA) to simplify the optimistic optimization. In this way, we convert an infinite-dimensional online learning problem into a decomposable and tractable one. We show that the proposed continuous-approximation optimistic (CA-O) learning algorithm can be implemented efficiently when the objective function exhibits special structures, such as concavity or convexity.

   c) For a broader scope of applications, maximizing over parameter uncertainty remains intricate due to potential nonconvexity or even the absence of a closed-form expression. Fortunately, owing to the continuous functions provided by CA, we propose an alternative algorithm named CA-O Faster Learning by applying first-order approximation with respect to parameters. We further derive a closed-form expression for the profit gradient used in the improved algorithm, regardless of whether a closed-form solution to the CA model exists. The closed-form gradient significantly improves precision and streamlines computation.

   d) We theoretically analyze the regret performance of both algorithms by examining two main gaps. The first is the CA gap, which occurs when translating a continuous solution into a discrete location design. The second is the learning gap incurred in the learning process due to a lack of knowledge of parameters. Coupling these two parts, we characterize the regret bound as comprising a moderate linear term from the CA gap and a sublinear term corresponding to the learning process.

III. *Numerical performance:* We test two algorithms with synthetic data. Both show sublinear regret, but CA-O Faster Learning is approximately 200 times more computationally efficient. We then apply the CA-O Faster Learning algorithm to a real-world case study in Toronto, where results indicate rapid convergence to a near-optimal store location design in just a few rounds. The CA-O Faster Learning algorithm helps retailers earn higher profits than benchmarks. The value of mobile retail stems from two aspects: Store mobility boosts net profit by 2.07% by addressing demand dynamics. Meanwhile, using the learning algorithm to overcome observational demand noise contributes extra profit. The value of learning increases from 3.29% to 13.03% of the overall profit in this case study as noise grows.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the online mobile retail store locations problem and discusses three main challenges. Section 4 proposes the online learning framework and two algorithms. Examples are provided to illustrate the application scope of different algorithms. Section 5 analyzes the CA gap and characterizes the regret bound of online learning algorithms. Numerical experiments are presented in Section 6, followed by the conclusion in Section 7. Detailed proofs are provided in the appendices.

2 Literature review

Dynamic facility location. Online facility location is increasingly relevant to an urban future with mobile facilities, but has drawn little attention in the literature. The majority of the dynamic facility location models are deterministic, assuming fully known information over the planning horizon. This stream of literature (e.g., Wesolowsky (1973), Canel et al. (2001), see Boloori and Zanjirani Farahani (2012) for a comprehensive review) typically focuses on customizing solution algorithms that combine dynamic programming with branch-and-bound or heuristic methods. On the other hand, studies on the stochastic problem in which customer demands vary as stochastic processes are relatively few.

However, neither deterministic nor stochastic dynamic facility location models are readily applicable to fast-evolving business environments, where the on-hand information is insufficient. Under such circumstances, the ability of online learning to make decisions while updating demand estimation becomes imperative. Despite its significance, the literature on facility location in online learning settings remains scarce. Bhatti et al. (2015) consider a two-stage location problem for planning alternative fuel stations with the ability to learn the demand and add more stations in the second stage. Nevertheless, it remains unknown how to constantly adjust locations in the presence of rich contextual data.

One stream of literature (Meyerson 2001) from the computer science community studies a variant online facility location problem where demands arrive sequentially. The decision is whether and where to build the next facility to minimize costs. However, facilities in their framework are irreversible, and locations do not influence exogenous demand. Subsequent studies (Guo et al. 2020, Kaplan et al. 2023) are confined to clustering and network design. In contrast, we focus on an learning-and-earning fashion, requiring a more flexible framework for rapidly changing urban business contexts. Moreover, our framework accounts for the interdependence between facility locations and unknown demands.

Continuous approximation. Our proposed CA-O learning algorithm employs continuous approximation (CA) to overcome the computational challenge associated with large-scale discrete facility location problems. The CA approach has been widely applied for various facility location problems. We refer readers to Ansari et al. (2018) for a recent survey. Among the papers that advance the CA method, Wang et al. (2017) propose a CA model to solve the dynamic facility location problem (yet with known parameters). Our paper makes methodological contributions to the CA literature by proposing an algorithmic framework to incorporate CA in an online-learning setup. Meanwhile, we show that the gap incurred by CA is moderate. Other contexts that employ CA include, e.g., the designs of supply chains (Lim et al. 2017), delivery system with drones (Carlsson and Song 2018), and retail store layout (Belavina 2021). Most recently, Blanchard et al. (2024) provide probabilistic approximations of k-traveling salesman problem and traveling repairman problem.

Combinatorial and continuous-armed bandits. Our paper advances the literature of bandits problems in both combinatorial and continuous-armed settings. When the candidate set of facility location is finite, online facility location degenerates into the area of combinatorial bandits. A combinatorial bandit is a linear bandit problem with action set that belongs to a d-dimensional binary hypercube (Cohen et al. 2017, Modaresi et al. 2020). In the online facility location problem, the total profit has a nonlinear structure, which much complicates the problem. If envisioning each possible combination as an arm, the problem is related to bandits with correlated rewards but only a few paper addresses this case (Ryzhov and Powell 2009, Ryzhov et al. 2012). When the possible facility locations lie in a continuous space, our problem is closely related to bandits with continuous actions. Exploring all arms is not feasible in bandits with continuous actions (non-combinatorial) (Agrawal 1995, Bubeck et al. 2011, Krishnamurthy et al. 2020). Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) study bandits problem with possibly infinite numbers of arms when expected rewards are linear functions of a scalar and a vector, respectively. More complicatedly, online facility location problem can be envisioned as a coupling of bandits with continuous actions and combinatorial bandits. The decision variable is a binary function on a multi-dimensional space where value 1 indicates the selection

of the facility location. It is challenging to solve the online problem with both low regret guarantee and low computational cost.

Decision-making with contextual information. For contextual bandits, upper-confidence bound (UCB) algorithms are a celebrated class of algorithms that are shown to have nice empirical performance (Bietti et al. 2021). A fair amount of works have been developed for linear bandits (Dani et al. 2008, Chu et al. 2011, Agrawal and Devanur 2019) and generalized linear models (GLM) (Li et al. 2017, Kveton et al. 2020). More recently, in the optimization community, there is an emerging interest in developing frameworks that integrate decision optimization and statistical model estimation (Ban and Rudin 2019, Bertsimas and Kallus 2020, Elmachtoub and Grigas 2022, Ho-Nguyen and Kılınç-Karzan 2022, Han et al. 2023). In our framework, both profit and response functions are parametric forms of contextual information. However, as the relationship is unknown, exploration is required to infer the true functions, through which to adaptively optimize decisions.

Urban retail and logistics. More broadly, our paper contributes to the growing literature on innovative urban retail and logistics. Examples of flexible retail stores include pop-up stores (Zhang et al. 2019), buy-online-pick-up-in-store fulfillment (Glaeser et al. 2019), and autonomous mobile vending stalls (Cao and Qi 2023). In a broader scope of logistics, there have been studies such as agile consolidation hubs (Wang et al. 2020), lockers (Lyu and Teo 2022), and urban aerial mobility (Kai et al. 2022). Our work complements these papers by theorizing the online location adjustment of flexible facilities.

3 The model of mobile retail stores with online location adjustment

This section models the sequential decision making for the mobile store location problem. We first introduce the problem formulation in Section 3.1, and then, analyze three main challenges in Section 3.2. In Section 3.3, we describe the technique to decide the store locations with known demand, i.e., the single period offline counterpart. A summary of notation is provided in Appendix A.

3.1 Formulation

Operations of mobile retail stores. Consider a retailer running a fleet of mobile retail stores to serve customers across an urban area $\mathcal{X}_t$ on day $t = 1, \cdots, T$. The customers naturally form Voronoi-shaped service zones centered at mobile retail stores as they go to the nearest store to make purchases. The retailer adjusts the locations of stores on a daily basis, with the objective of finding the optimal store location design to maximize profit by selling more products and saving costs.

In the dynamic environment of mobile retail, exact demand locations are numerous and difficult to identify. Mobile retail stores are small-scale and flexible in location decisions. Given the vast number of potential store and demand locations, determining the exact locations becomes impractical. Thus, instead of formulating conventional mixed-integer programs for location problems, we consider a continuous service area, $\mathcal{X}_t$, which is also the set of candidate store locations. The decisions are to dynamically adjust a set of N_t store locations $\boldsymbol{x}_t = \{x_{t1}, x_{t2}, ..., x_{tN_t}\}$ over time t, such that $A_t(x) = 1$ for $x \in \boldsymbol{x}_t$ and $A_t(x) = 0$ otherwise, where $A_t \in \mathcal{A}_t$ and $\mathcal{A}_t$ is the set of all feasible actions on day t. The choice of store locations automatically partitions space $\mathcal{X}_t$ into a set of non-overlapping *influence areas* (i.e., service zones), $\boldsymbol{\mathcal{X}}_t = \{\mathcal{X}_{t1}, \mathcal{X}_{t2}, ..., \mathcal{X}_{tN_t}\}$, such that $\boldsymbol{\mathcal{X}}_t = \bigcup_i \mathcal{X}_{ti}$ and $\mathcal{X}_{ti} \cap \mathcal{X}_{tj} = \emptyset$ for $i \neq j$. Since the decision A_t and $(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t)$ have one-to-one mapping, we can rewrite A_t as $A_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t)$, i.e., $A_t(x) = A_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t)(x)$ for all $x \in \mathcal{X}$.

The influence areas represent one core trade-off in the decision-making for A_t. On one hand, when setting larger influence areas, the retailer lowers operating costs by deploying fewer mobile retail stores. However, the costs of replenishment per unit increase since trucks travel longer distances to restock each store. Larger influence areas also result in disutility for customers, because customers have to travel farther to visit these stores, which in turn reduces the retailer's revenues.

Demand function. We model the customer demand with contextual information. At the beginning of day t, the retailer observes a context function $W_t(x) : \mathcal{X}_t \to \mathcal{W}_t$. We assume that demand locations are distributed according to a continuous spatial density function, denoted by $\rho_{\theta^*}(A_t, x; W_t(x))$ (per day per kilometer squared) for $x \in \mathcal{X}$. This demand density function is parameterized by θ^*, and also depends on the action $A_t \in \mathcal{A}_t$ and local context $W_t(x)$. We specifically assume that a kernel vector $\kappa(A_t, W_t(x)) \in \mathbb{R}^d$ describes the features at location $x \in \mathcal{X}$ such that

$$\rho_\theta(A_t, x; W_t(x)) = \theta^\top \kappa(A_t, W_t(x)). \tag{1}$$

The features can be, for example, local population, distance to the store, traffic condition, etc. In Section 6, we also provide a thorough discussion of the features we used in the case study in Toronto. Our setting is general to allow the kernel function κ to potentially change over time t.

The profit of mobile retail stores depends on customer demand, but the retailer is unaware of the demand because the parameter θ^* is unknown and can only be estimated from historical observations. However, the exact value of $\rho_{\theta^*}(A_t, x; W_t(x))$ at $x \in \mathcal{X}_t$ is inaccessible, since demand is realized at each store at location x_{tj} rather than every point $x \in \mathcal{X}_t$ over the entire area. We assume Y_{tj} is the demand served by the store at location x_{tj} on day t, such that its relationship with the explanatory variables (A_t, W_t) is as follows:

$$Y_{tj} = f_{\theta^*}(A_t; W_t, \mathcal{X}_{tj}) + \epsilon_{tj},$$

where

$$f_\theta(A_t; W_t, \mathcal{X}_{tj}) = \int_{\mathcal{X}_{tj}} \rho_\theta(A_t, x; W_t(x))dx.$$

Moreover, let

$$\mathcal{H}_t := \sigma(A_1, W_1, Y_1, \cdots, W_{t-1}, A_{t-1}, Y_{t-1}, A_t, W_t)$$

be the σ-algebra summarizing the information available just before observing the response $\mathbf{Y}_t := \{Y_{tj}; j = 1, ..., N\}$. We assume that the observational noise ϵ_{tj} is $\mathcal{H}_t$-measurable and $\mathbb{E}[\epsilon_{tj}|\mathcal{H}_t] = 0$.

Objectives. The retailer's objective is to find a sequence of mobile retail store location decisions to maximize the total expected profit. In other words, the retailer aims to solve a sequential problem:

$$\max_{\{A_t \in \mathcal{A}_t; t=1,...,T\}} \sum_{t=1}^{T} r_{\theta^*}(A_t; W_t), \tag{2}$$

in which $r_{\theta^*}(A_t; W_t) := \mathbb{E}[R_{\theta^*}|A_t, W_t]$ denotes the conditional expected profit of day t, and $\theta^* \in \mathbb{R}^d$ is the unknown parameter vector. On each day t, the retailer observes the context $W_t(x)$, chooses a store location action $A_t \in \mathcal{A}_t$, observes the response Y_t, and receives a profit R_t. The fundamental problem in this paper is to simultaneously explore (to estimate θ^*) and exploit (to maximize profit) over time. Through exploration, the retailer consciously sacrifices immediate profits in exchange for valuable demand information, which empowers the retailer to make better decisions and consequently secure higher future profits. However, if the retailer commits exclusively to exploiting current information for actions, they run the risk of being blind to the demand variations in certain regions of area $\mathcal{X}_t$ or certain dimensions of θ^*. This oversight leads to missed prospects for long-term profit maximization.

Define $\varphi_{\theta^*}(A_t, x; W_t(x))$ as the per-km^2 expected profit of serving demands around location x via a store deployed. The profit equals the revenue from selling products, minus the cost of inventory replenishment, and the operating cost of mobile retail stores, which we also call facility cost in our notation. More specifically, the functional form can be expressed as follows:

$$\varphi_\theta(A_t, x; W_t(x)) = \underbrace{\bar{r}\rho_\theta(A_t, x; W_t(x))}_{\text{Revenue density}} - \underbrace{\varphi^i\left(\int_{\mathcal{X}_{tj}} \rho_\theta(A_t, x; W_t(x))dx, \mathcal{X}_{tj}; W_t(x)\right)}_{\text{Inventory replenishment cost density}} - \underbrace{\varphi^f\left(\int_{\mathcal{X}_{tj}} \rho_\theta(A_t, x; W_t(x))dx, \mathcal{X}_{tj}; W_t(x)\right)}_{\text{Facility cost density}}, \tag{3}$$

in which we define $\bar{r}$ as the average revenue per customer, $d(x_{tj}, x)$ as the distance between a store at x_{tj} and a customer at x within its influence area. $\int_{\mathcal{X}_{tj}} \rho_\theta(A_t, x; W_t(x))dx$ is the expected daily sales handled by the store serving $\mathcal{X}_{tj}$. The inventory replenishment cost involves transporting goods from a warehouse to multiple stores via truck routing. The facility cost includes fixed opening cost, and goods handling cost (which is proportional to the daily sales). We omit the cost of repositioning stores across days from A_t to A_{t+1}, as the retailer dispatches mobile stores to their bases at the end of each day t.

We include $\int_{\mathcal{X}_{tj}} \rho_\theta(A_t, x; W_t(x))dx$ as an input in the inventory replenishment and the facility cost density functions to emphasize that these two cost densities are calculated at the influence-area level and then evenly allocated to each $x \in \mathcal{X}_t$. We would also like to emphasize that the profit density $\varphi_\theta(A_t, x; W_t(x))$ at $x \in \mathcal{X}$, by construction, depends not only on local action $A_t(x)$ and local covariates $W_t(x)$, but also on $A_t(x')$ for $x' \neq x$, due to the combinatorial nature of the problem.

Under such a setting, the expected profit at day t is

$$r_{\theta^*}(A_t; W_t) = \sum_{j=1}^{N_t} \left(\int_{x \in \mathcal{X}_{tj}} \varphi_{\theta^*}(A_t, x; W_t(x))dx\right),$$

and the overall problem (2) can be more explicitly rewritten as the following online facility location (OFL) problem:

$$\max_{\{(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t); t=1,\ldots,T\}} \sum_{t=1}^{T} \sum_{j=1}^{N_t} \left(\int_{x \in \mathcal{X}_{tj}} \varphi_{\theta^*}(A_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t), x; W_t(x))dx\right). \tag{OFL}$$

3.2 Challenges

Solving the online facility location problem (OFL) is nontrivial. We identify three main challenges:

Complexity of action space: Most papers in the existing bandits literature assume the action set $\mathcal{A}$ to be a space of variables. In contrast, in the online facility location problem, $\mathcal{A}$ is instead a space of functions over a multi-dimensional space. That is, each of its element $A(x)$ is defined on a continuous domain $\mathcal{X}$. The problem of selecting an optimal function is much more complicated than choosing a variable, especially in an online-learning setting in which exploration and exploitation need to be balanced. Even if we instead assume $\mathcal{X}$ to be discrete and finite so that the problem falls into the scope of combinatorial bandits, the problem is still much more challenging than what existing generic algorithms can handle. This is because the total profit function in location problems may often involve nonlinear structures (e.g., inventory costs and routing costs) that couple individual costs. Alternatively, if treating each combination as an independent arm, the regret would increase exponentially with the cardinality of $\mathcal{X}$. Therefore, we need to leverage analytical convenience that is inherent in facility location models to design a learning algorithm over an action space with high complexity (either a functional space or combinatorial action space).

Computational intractability from optimistic algorithms: Computational tractability is another challenge in designing the learning algorithm. The infinite action space usually incurs

computational hurdles, even when optimizing over a variable rather than a function. In particular, the UCB algorithm constructs an uncertainty set for the parameter and solves a max-max problem over the joint parameter and action set. However, solving the optimistic optimization (max-max) problem for large or continuous action sets is often intractable due to the potential nonconvexity of the problem. Even in a simple scenario of linear bandits with infinite actions, solving the max-max problem entails a bilinear optimization problem. A similar issue also exists in our (OFL) setup, where the domain $\mathcal{X}$ is continuous, and infinitely many choices exist since the binary function space contains an infinite number of functions. Given the learning complexity and optimization complexity, it is vital to design a computationally efficient online algorithm with low-regret guarantees.

Regret analysis: Finally, we need to quantify the performance of the proposed optimal learning algorithms. Since these new algorithms are customized for tackling the first two challenges, we cannot directly borrow existing approaches, but have to conduct new analysis of the regret benchmarked against the offline, full-information baseline.

3.3 Continuous approximation and cost analysis

We address the first challenge by simplifying the action space using a continuous approximation (CA) approach. Meanwhile, we apply the CA approach to provide an estimation of the costs incurred in the operations of mobile retail stores.

We start with an offline, single-period, static formulation, in which the retailer only considers a one-shot optimization problem to maximize the profit function with complete information on the demand (when the true parameter θ^* is known). Even so, the store location model described in Section 3.1 is generally difficult to solve. As discussed in the first challenge in Section 3.2, the action space $\mathcal{A}$ is infinite. Even if the action space is finite, enumerating all possibilities is likely to be computationally infeasible. In addition, the profit density function $\varphi(\cdot)$ involves norms such as $\|x_{tj} - x\|$ to account for the distance from a store to a point within its influence area $\mathcal{X}_{tj}$. It is inconvenient to directly use integrals of such norm functions to optimize discrete facility locations and partition the service zone.

To overcome these obstacles, we utilize a CA approach. The main idea of CA is that the size of influence areas $\mathcal{X}_{tj}$ can be approximated by a continuous *influence area function* $z_t(x)$ for $x \in \mathcal{X}_{tj}$, i.e., $|\mathcal{X}_{tj}| \approx z_t(x)$ where $z_t \in \mathcal{Z}_t$. The set $\mathcal{Z}_t$ is a class of non-negative and continuous functions over $\mathcal{X}_t$. The decision of the CA problem is $z_t(x)$ instead of the binary action function $A_t(x)$. This approximation has been extensively tested to result in small errors in approximating the optimal objective value if $z_t(x)$ is slow-varying in x and if the influence areas are near "round" with stores located near their centers, which are the case in the operations of mobile retail stores and indeed the case in near-optimal designs under mild parameter conditions (Daganzo 2005). Subsequently, the profit density function $\varphi_\theta(A_t(\boldsymbol{x}, \boldsymbol{\mathcal{X}}); W_t(x))$ in (OFL) can be approximated by a continuous function $\psi_\theta(z_t(x); W_t(x))$, yielding the following single-period CA model:

$$\max_{A_t \in \mathcal{A}_t} \int_{x \in \mathcal{X}_t} \varphi_\theta(A_t, x; W_t(x)) dx \quad \approx \quad \max_{z_t \in \mathcal{Z}_t} \int_{x \in \mathcal{X}_t} \psi_\theta(z_t(x); W_t(x)) dx. \tag{4}$$

The advantage of CA formulation on the right-hand side of (4) is that it can be *decomposed* and then efficiently optimized with respect to each location x by finding the optimal solution $z_t^*(x)$ that maximizes the integrand $\psi_\theta(z_t(x); W_t(x))$. This is because, whereas $\varphi_\theta(A_t, x; W_t(x))$ depends on the function $A_t(x')$ for $x' \in \mathcal{X}$ and $x' \neq x$, $\psi_\theta(z_t(x); W_t(x))$ only depends on value $z_t(x)$ locally at x. Specifically,

$$z_t^*(x; \theta) = \arg\max_{z' \in \mathbb{R}} \ \psi_\theta(z'; W_t(x)). \tag{5}$$

The continuous profit density function $\psi_\theta(\cdot)$ follows the similar structure as in (3) (for brevity, we suppress the dependence of ρ and z on other quantities such as θ, W and t wherever appropriate):

$$\psi_\theta(z(x);W(x)) = \underbrace{\bar{r}\rho(x)}_{\text{Revenue}} - \underbrace{\varphi^i\Big(\rho(x)z(x),z(x);W(x)\Big)}_{\text{Inventory replenishment cost}} - \underbrace{\varphi^f\Big(\rho(x)z(x),z(x);W(x)\Big)}_{\text{Facility costs}}. \tag{6}$$

These three terms are obtained through approximations: the store influence area $|\mathcal{X}_{tj}|$ is represented by $z(x)$ and daily sales $\int_{\mathcal{X}_{tj}}\rho(x)dx$ by $\rho(x)z(x)$. The demand density $\rho(x)$ is short for $\rho_\theta(z(x),x;W(x))$, which depends on influence area $z(x)$ and feature $W(x)$ around location x. The inventory replenishment cost and facility cost are given by

$$\varphi^i\Big(\rho(x)z(x),z(x)\Big) = \frac{\rho(x)z(x)}{S}\cdot\frac{\beta_{\mathsf{TSP}}c_t}{\sqrt{z(x)}} = \beta_{\mathsf{TSP}}\frac{c_t}{S}\rho(x)\sqrt{z(x)}, \tag{7a}$$

$$\varphi^f\Big(\rho(x)z(x),z(x)\Big) = \frac{a^f\rho(x)z(x)+b^f}{z(x)}. \tag{7b}$$

Here $S, c_t, \beta_{\mathsf{TSP}}, a^f, b^f$ are cost parameters. We first quantify the daily truck routing costs for inventory replenishment. Since a truck visits multiple stores per trip, the routing costs also depend on other nearby store locations. Fortunately, we can approximate the replenishment frequency locally using CA. For any store at a location $x\in\mathcal{X}_t$, we determine the average number of daily replenishments so that the volume of each refill, denoted as S, is a specific portion of the store's capacity. Recall that the average daily sales of the store is $\rho(x)z(x)$. Therefore, we obtain the replenishment frequency as $\rho(x)z(x)/S$. Suppose that the truck incurs a cost of c_t per kilometer. The routing distance is obtained from the traveling salesman problem (TSP) under Euclidean metric. The well-known BHH Theorem (Beardwood et al. 1959) provides an approximation of the optimal TSP tour as $\beta_{\mathsf{TSP}}\int_{x\in\mathcal{X}_t}1/\sqrt{z(x)}dx$, where β_{TSP} is a constant; we use the estimation $\beta_{\mathsf{TSP}}\approx 0.7124$, as suggested in Applegate et al. (2010). Thus, the cost density of one trip is $\beta_{\mathsf{TSP}}c_t/\sqrt{z(x)}$. Multiplying the frequency by one trip routing cost immediately yields the estimation of replenishment cost density $\varphi^i(\cdot)$ in (7a). Afterward we estimate the facility costs by denoting the goods handling costs as $a^f\rho(x)z(x)$ and the fixed opening cost of a store as b^f. Since the facility costs are incurred by a store covering area $z(x)$, one can obtain the facility cost density $\varphi^f(\cdot)$ in (7b). We will analyze three cases within this basic setting in Section 4, including concave, convex functions, and functions lacking a closed-form maximizer.

Once obtaining the optimal solution $z^*(\cdot;\theta)$ with parameter θ, one can translate the CA recipe into discrete store location decisions, denoted by $A(z^*(\cdot;\theta))$, by applying a discretization procedure. Then the final profit of action $A(z^*(\cdot;\theta))$ is

$$r_\theta(A(z^*(\cdot;\theta));W) = \int_{x\in\mathcal{X}}\varphi_\theta(A(z^*(\cdot;\theta)),x;W(x))dx \approx \int_{x\in\mathcal{X}}\psi_\theta(z^*(x;\theta);W(x))dx.$$

For notation brevity, hereafter we define the *approximate objective function* from the CA model (4) as

$$r^\psi_\theta(z;W) := \int_{x\in\mathcal{X}}\psi_\theta(z(x);W(x))dx. \tag{8}$$

We will analyze the error induced by CA in Section 5.

Remark. The functional form (7) is specific to the mobile retail problem. In Appendix E, we analyze additional settings beyond the scope of mobile retail stores as model extensions to enhance the general applicability of our model, such as one-to-one inventory replenishment, delivery products to customers, and last-mile delivery using micro-depots.

4 The CA-O learning algorithm

We now proceed to develop learning algorithms that address the operations of mobile retail stores over the planning horizon T, with the retailer seeking to maximize overall profits. In Section 4.1 we

propose an algorithmic framework for solving the sequential decision making problem (OFL). Moving on to Section 4.2, we tackle the second challenge mentioned in Section 3.2 by designing an alternative algorithm that is computationally efficient. We will address the third challenge of quantifying regret in Section 5.

4.1 A learning framework

Having described the CA technique for the single-period problem, we move on to the online and multi-period setting with parameter learning incorporated. A general principle of such decision making is optimism in the face of uncertainty. This principle is particularly embodied by the UCB algorithm, which has been applied to a wide range of optimization problems. The benefit of the UCB algorithm in mobile retail store operations is its ability to achieve balance between maximizing profits and gathering information about demand across various service regions and dimensions of θ^*, all while optimizing actions efficiently over the time horizon.

The key step in the UCB algorithm is to construct a confidence set $\Theta_t \subset \mathbb{R}^d$ based on $\mathcal{H}_t$. Similar to linear bandits, there are conflicting desirable properties for constructing Θ_t: Θ_t should contain the unknown parameter θ^* with high probability and Θ_t should be as small as possible. When Θ_t contains the true parameter θ^*, $\max_{\theta\in\Theta_t} \max_{A\in\mathcal{A}_t} r_\theta(A; W_t)$ provides an upper bound for the true optimal objective value. For a given action $A \in \mathcal{A}_t$ and confidence set Θ_t, let

$$\mathsf{UCB}_t(A) = \max_{\theta\in\Theta_t} \; r_\theta(A; W_t)$$

be an upper-confidence-bound of the expected reward of action A, and the reward in our (OFL) setup is the overall profit of a mobile retail store location design. Therefore, $\mathsf{UCB}_t(A)$ is an optimistic estimator. The UCB algorithm selects action A_t at time t such that

$$A_t = \arg\max_{A\in\mathcal{A}_t} \; \mathsf{UCB}_t(A) = \arg\max_{A\in\mathcal{A}_t} \max_{\theta\in\Theta_t} \; r_\theta(A; W_t). \tag{9}$$

To solve this problem, we propose a *Continuous-Approximation Optimistic Learning* (CA-O Learning.) Algorithm. The idea is to combine the CA technique with a new UCB algorithm. Specifically, having simplified the action space to locationwise-decomposable influence area functions, the first step is to simply replace problem (9) with the following CA problem

$$\max_{\theta\in\Theta_t} \max_{z\in\mathcal{Z}_t} \; r^\psi_\theta(z; W_t) \tag{10}$$

to reduce the complexity of the action space. Since $z^*_t(x;\theta)$ for any given θ can be efficiently evaluated point by point, we can rewrite the optimization problem (10) as follows:

$$\max_{\theta\in\Theta_t} \; r^\psi_\theta(z^*_t; W_t) = \max_{\theta\in\Theta_t} \max_{z\in\mathcal{Z}_t} \; r^\psi_\theta(z; W_t) \overset{(5)}{=} \max_{\theta\in\Theta_t} \int_{x\in\mathcal{X}} \psi_\theta(z^*_t(x;\theta); W_t(x))dx. \tag{OFL-CA}$$

The next step is to construct the uncertainty set Θ_t in (OFL-CA). Each historical observation is represented by a triple $(\mathbf{Y}_t, W_t, A_t)$ where $\mathbf{Y}_t \in \mathbb{R}^{N_t}$ is a vector of responses at time t. Given $t-1$ observations, $\{(\mathbf{Y}_s, W_s, A_s)\}_{s=1}^{t-1}$, we suppose that the model parameter θ^* can be estimated by minimizing a statistical squared-loss function on $\ell^\lambda_t : \mathbb{R}^d \to \mathbb{R}$:

$$\hat{\theta}_t \in \arg\min_{\theta\in\mathbb{R}^d} \; \ell^\lambda_t(\theta) = \arg\min_{\theta\in\mathbb{R}^d} \; \sum_{s=1}^{t-1} \sum_{j=1}^{N_s} (Y_{sj} - f_\theta(A_s; W_s, \mathcal{X}_{sj}))^2 + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$. We use the shorthand $f_{sj}(\theta) := f_\theta(A_s; W_s, \mathcal{X}_{sj})$ to denote the mean demand at the influence area $\mathcal{X}_{sj}$. Now consider a supervised learning oracle that outputs a root of the following

equation of the gradient of the loss function:

$$\nabla_\theta \ell_t^\lambda(\hat{\theta}_t) = \sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - Y_{sj})g_{sj} + \lambda\hat{\theta}_t = 0, \tag{Oracle}$$

where $g_{sj} = \nabla_\theta f_{sj}(\theta) = \int_{\mathcal{X}_{sj}} \kappa(A_s, W_s(x))dx$. (Oracle) can be viewed as the first-order condition for minimizing the loss function. For a fixed λ, define the design matrix

$$V_t = \sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}{g_{sj}}^\top + \lambda I. \tag{11}$$

The matrix V_t is determined by historical actions and contexts. V_t plays an important role in constructing the uncertainty set. At time step t, we define the uncertainty set as

$$\Theta_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t} \le \gamma_t\},$$

which is an ellipsoid centred at $\hat{\theta}_t$ with principal axes being the eigenvectors of V_t and the radius being γ_t. The corresponding eigenvalues of V_t are increasing with time, so the radius of the ellipse is decreasing. With a proper choice of γ_t, the designed algorithm guarantees that the true parameter θ^* is contained in Θ_t with high probability.

Based on the above two steps, we outline CA-O Learning. Algorithm to solve Problem (OFL). In line 1-3, we randomly explore actions in the first t_0 time periods. At time $t = t_0 + 1, \cdots, T$, we first compute an estimator $\hat{\theta}_t$ by solving (Oracle). Then we construct an elliptical uncertainty set Θ_t in line 7. In line 8, we first solve problem (5) for the analytical expression of the size of the continuously approximated influence area $z_t^*(x;\theta)$ as a function of θ, and then solve the optimistic optimization problem (OFL-CA) for the optimistic parameter estimator θ_t over Θ_t. Finally, in line 9, we apply a discretization procedure to functions $z_t^*(\cdot;\theta_t)$ and then implement location decisions $A(z_t^*(\cdot;\theta_t))$.

Algorithm CA-O Learning.

Input: time horizon length T and exploration periods t_0
1: **for** $t = 1, \cdots, t_0$ **do**
2: Choose decision $A_t \in \mathcal{A}_t$ according to the sampling rule, and receive response $\mathbf{Y}_t$;
3: **end for**
4: **for** $t = t_0 + 1, \cdots, T$ **do**
5: Compute $\hat{\theta}_t$ by solving (Oracle);
6: Update V_t according to Equation (11);
7: Update $\Theta_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t} \le \gamma_t\}$;
8: Derive the analytical expression $z_t^*(\cdot;\theta)$ from problem (5), and then solve (OFL-CA) for the optimistic estimator θ_t;
9: Implement $A_t = A_t(z_t^*(\cdot;\theta_t))$ based on continuous solution $z_t^*(\cdot;\theta_t)$; Receive response Y_t;
10: **end for**

The following case, a slight variant of the cost formulas specified in (7), illustrates how to apply CA-O Learning. Algorithm. We consider the case where the objective function of (OFL-CA) is concave in θ, so that the step (line 8) of solving (OFL-CA) can be efficient.

Case 1 (Concave Objective Function)**.** In the initial stage of mobile retail store deployment, the retailer has the option to utilize crowdsourcing delivery to restock their stores. Crowdsourcing offers retailers an asset-light strategy for establishing their logistics. However, the cost of crowdsourcing might increase with higher demand, as the drivers bid on tasks with fluctuating prices. Based on cost formulas (7), we further assume the unit travel cost of replenishment increases with demand, i.e., $c_t = \bar{c}_t\rho_\theta(x)$, in which $\bar{c}_t$ is a constant. We also consider the demand kernel function $\kappa(A_t, W_t(x)) = W_t(x)$, so that the demand density only depends on contextual covariates and $\rho_\theta(x) = \theta^\top W_t(x)$. The CA of profit function is

$$r_\theta^\psi(z_t; W_t) = \int_{x\in\mathcal{X}_t} \psi_\theta(z_t; W_t(x))dx = \int_{x\in\mathcal{X}_t} \left(\bar{r}\rho_\theta(x) - \beta_{\mathsf{TSP}}\frac{\bar{c}_t}{S}\rho_\theta^2(x)\sqrt{z_t(x)} - \frac{a^f\rho_\theta(x)z_t(x) + b^f}{z_t(x)}\right)dx.$$

Applying the first-order condition to this CA model yields the following optimal solution and optimal profit density function, respectively:

$$z_t^*(x;\theta) = \left(\frac{2b^f S}{\beta_{\mathsf{TSP}}\bar{c}_t\rho_\theta^2(x)}\right)^{\frac{2}{3}},$$

$$\psi_\theta(z_t^*(x;\theta);W_t(x)) = (\bar{r}-a^f)\rho_\theta(x) - 3\left(b^f\right)^{\frac{1}{3}}\left(\frac{\beta_{\mathsf{TSP}}\bar{c}_t}{2S}\right)^{\frac{2}{3}}\rho_\theta^{\frac{4}{3}}(x).$$

Given that $\rho_\theta(\cdot)$ is a linear function in θ, $\psi_\theta(z_t^*(x;\theta);W_t(x))$ is concave in θ. It follows that the objective of (OFL-CA), $r_\theta^\psi(z_t^*(\cdot;\theta);W_t)$, is also concave in θ, as $r_\theta^\psi(z_t^*(\cdot;\theta);W_t)$ is an integral of $\psi_\theta(z_t^*(x;\theta);W_t(x))$ over $x\in\mathcal{X}$. Additionally, since Θ_t is a convex set, the maximization problem (OFL-CA) becomes a tractable convex optimization problem with a differentiable objective function. This problem can be efficiently solved using convex optimization algorithms. ♣

As one of the main algorithms proposed in this paper, CA-O Learning. resolves the complexity of the action space by utilizing the structural convenience of CA, embeds a UCB-type of strategy to balance the exploration vs. exploitation trade-off, and invokes an influence-area-discretization recipe. Before jumping into the regret analysis for this algorithm, we need to overcome one more obstacle: Echoing the second challenge stated in Section 3.2, the optimization problem (OFL-CA) in line 8 may be computationally difficult. We solve this issue and alternatively propose Algorithm CA-O Faster Learning in the next subsection.

4.2 Computational challenges and CA-O faster learning

We first discuss when the optimization problem (OFL-CA) is readily solvable. As demonstrated in Case 1, when the function $r_\theta^\psi(z^*(\cdot;\theta);W)$ is concave in θ, the maximization over a convex set can be addressed using the first-order condition. Conversely, if $r_\theta^\psi(z^*(\cdot;\theta);W)$ is convex in θ, optimizing it may result in reduced computational efficiency, as illustrated in the subsequent case.

Case 2 (Convex Objective Function)**.** We keep the same assumption as in Case 1 that the demand density function is represented by $\rho_\theta(x)=\theta^\top W_t(x)$. We alternatively consider the basic operational setting of mobile retail stores, with cost formulas as given by (7). The CA of profit function is expressed as

$$r_\theta^\psi(z_t;W_t) = \int_{x\in\mathcal{X}_t}\psi_\theta(z_t;W_t(x))dx = \int_{x\in\mathcal{X}_t}\left(\bar{r}\rho_\theta(x) - \beta_{\mathsf{TSP}}\frac{c_t}{S}\rho_\theta(x)\sqrt{z_t(x)} - \frac{a^f\rho_\theta(x)z_t(x)+b^f}{z_t(x)}\right)dx.$$

At each x, the optimal solution $z_t^*(x;\theta)$ and the optimal profit density are given by

$$z_t^*(x;\theta) = \left(\frac{2b^f S}{\beta_{\mathsf{TSP}}c_t\rho_\theta(x)}\right)^{\frac{2}{3}},$$

$$\psi_\theta(z_t^*(x;\theta);W_t(x)) = (\bar{r}-a^f)\rho_\theta(x) - 3\left(b^f\right)^{\frac{1}{3}}\left(\frac{\beta_{\mathsf{TSP}}c_t}{2S}\rho_\theta(x)\right)^{\frac{2}{3}},$$

which indicates that $\psi_\theta(z_t^*(x;\theta);W_t(x))$ is convex in θ, and it follows that $r_\theta^\psi(z_t^*(\cdot;\theta);W_t)$ is convex in θ. Thus, if Θ_t is a convex hull of a finite set, it suffices to enumerate values of θ over a finite set of extreme points, but the number of enumerations can be large. Especially, we construct Θ_t as an ellipsoid set in this paper, which means that there are infinite many extreme points. As a result, the maximization problem (OFL-CA) becomes intractable in this case. ♣

In general, directly solving problem (OFL-CA) can be computationally cumbersome due to nonconvexity. To overcome this difficulty, we propose Algorithm CA-O Faster Learning. The idea is to change

Algorithm CA-O Faster Learning

Input: time horizon length T and exploration periods t_0^F
1: **for** $t = 1, \cdots, t_0^F$ **do**
2: Choose decision $A_t \in \mathcal{A}_t$ according to the sampling rule, and receive response $\mathbf{Y}_t$;
3: **end for**
4: **for** $t = t_0 + 1, \cdots, T$ **do**
5: Same as lines 5-7 in CA-O Learning.;
6: Compute $z_t^*(\cdot; \hat{\theta}_t)$ according to Equation (5) and compute θ_t by solving:

$$\theta_t = \arg\max_{\theta \in \Theta_t} \; r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t)^\top (\theta - \hat{\theta}_t);$$

7: Implement $A_t = A_t(z_t^*(\cdot; \theta_t))$ based on the continuous solution $z_t^*(\cdot; \theta_t)$; Receive response Y_t;
8: **end for**

line 8 of CA-O Learning.. Specifically, to optimize over θ, we instead use the first-order approximation as the objective function

$$r^{\psi}_{\theta}(z_t^*(\cdot; \theta); W_t) \approx r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t)^\top (\theta - \hat{\theta}_t),$$

where $\nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t)$ is the gradient of the composite function $r^{\psi}_{\theta}(z_t^*(\cdot; \theta); W_t)$ with respect to θ at $\hat{\theta}_t$. Under this approximation, we only need to find $\theta \in \Theta_t$ such that

$$\theta_t = \arg\max_{\theta \in \Theta_t} \; \nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot; \hat{\theta}_t); W_t)^\top (\theta - \hat{\theta}_t). \tag{12}$$

which has a closed-form solution at each day t as the following lemma shows.

Lemma 1. The optimal solution to Equation (12) is $\theta_t = \hat{\theta}_t + \gamma_t \frac{V_t^{-1} \nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot;\hat{\theta}_t);W_t)}{\|\nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot;\hat{\theta}_t);W_t)\|_{V_t^{-1}}}$.

This is a one-step computation after obtaining the estimator $\hat{\theta}_t$, and thus significantly improves the computational efficiency. Furthermore, there exists a scenario where CA-O Learning. fails and only CA-O Faster Learning can be applied - specifically when z_t^* cannot be solved analytically. In such instances, the optimal objective function ψ_θ does not even have a closed-form expression, as the following case illustrates.

Case 3 (No Closed-form Solution)**.** When store influence areas are large, the travel disutility for customers cannot be overlooked. Consequently, customer demand decreases as the distance to the stores increases. Following the customer demand model proposed by Berman et al. (1995), we assume that demand decays exponentially with distance. Specifically, for customers located at $x \in \mathcal{X}_t$, suppose the nearest store is situated at x_{tj}. i.e., these customers are within the influence area of store j. The demand density function can be expressed as

$$\rho_\theta(A_t, x; W_t(x)) = \theta^\top \kappa(A_t, W_t(x)) = \theta^\top W_t(x) \exp\left\{-c_0 d(x_{tj}, x)\right\},$$

where c_0 represents a constant parameter, and $d(x_{tj}, x)$ denotes the distance between customers at location x and the nearest store.

The CA approach yields the average demand density over a store's influence area $z(x)$, given as $\rho_\theta(z_t(x), x; W_t(x)) = \theta^\top W_t(x) \exp\left\{-c_0 \frac{2}{3\sqrt{\pi}} \sqrt{z_t(x)}\right\}$. In the logistics setting where cost formulas are defined by (7), the resulting profit density function is as follows:

$$\psi_\theta(z_t(x); W_t(x)) = \left(\bar{r} - a^f - \beta_{\mathsf{TSP}} \frac{c_t}{S} \sqrt{z_t(x)}\right) \theta^\top W_t(x) \exp\left\{-c_0 \frac{2}{3\sqrt{\pi}} \sqrt{z_t(x)}\right\} - \frac{b^f}{z_t(x)}. \tag{13}$$

Although we can numerically evaluate $z_t^*(x; \theta)$ by solving $\frac{\partial \psi_\theta}{\partial z_t}(z_t^*; W_t(x)) = 0$, a closed-form maximizer $z_t^*(x; \theta)$ does not exist. The resulting integral function $r_\theta^*(z_t^*; W_t)$ is thus implicit in the expression. ♣

In Case 3, (OFL-CA) cannot be reduced to an optimization problem solely with respect to θ in a closed-form expression, rendering CA-O Learning. inapplicable. The next step is to examine whether we can apply CA-O Faster Learning by computing $\nabla r^{\psi}_{\hat{\theta}_t}(z^*_t; W_t)$ in line 6. As aforementioned, we may not be able to obtain a closed-form solution for z^*_t, which $\nabla r^{\psi}_{\hat{\theta}_t}(z^*_t; W_t)$ depends on. In such a scenario, it becomes impossible to obtain a closed-form formula of $\psi_\theta(z^*_t(x;\theta); W_t(x))$ with respect to θ. The numerical computation of $\nabla r^{\psi}_{\theta}(z^*; W) = \int_{x\in\mathcal{X}} \nabla_\theta \psi_\theta(z^*(x;\theta); W(x))dx$ is required. However, the numerical differentiation $\nabla_\theta \psi_\theta(z^*(x;\theta); W(x))$ presents two issues. First, finite differences method is potentially ill-conditioned for the implicit function. Second, as the calculation of $\nabla_\theta \psi_\theta(z^*(x;\theta); W(x))$ involves $\nabla_\theta z^*(\cdot;\theta)$, evaluating $z^*(\cdot;\theta + d\theta)$ numerically introduces additional precision errors. Both issues worsen the error in our approximation algorithm.

Nevertheless, fortunately and surprisingly, the above potential issues can be avoided by Lemma 2, which provides an explicit and analytical formula for the gradient $\nabla_\theta \psi_\theta(z^*(x;\theta); W(x))$, without the knowledge of $\nabla_\theta z^*(\cdot;\theta)$. The value of Lemma 2 lies in its ability to avoid the complex computation of $\nabla_\theta z^*(\cdot;\theta)$.

Lemma 2. For any $\theta \in \Theta$ and $x \in \mathcal{X}$, we can compute the gradient of ψ_θ as follows

$$\nabla_\theta \psi_\theta(z^*(x;\theta); W(x)) = \left[\bar{r} - z^*(x;\theta)\left(\frac{\partial \varphi^i}{\partial(\rho(x)z^*(x;\theta))} + \frac{\partial \varphi^f}{\partial(\rho(x)z^*(x;\theta))}\right)\right] \kappa(z^*(x;\theta), W(x)).$$

In summary, if a closed-form solution $z^*_t(\cdot;\theta)$ exists and r^{ψ}_{θ} is concave in a maximization problem, Algorithm CA-O Learning. operates efficiently. If $z^*_t(\cdot;\theta)$ is in closed-form with convex r^{ψ}_{θ}, one can opt for CA-O Learning. and enumerate all extreme points of Θ_t if extreme points are of small size; otherwise, CA-O Faster Learning is the better option. If there is no closed-form solution $z^*_t(\cdot;\theta)$, Algorithm CA-O Faster Learning can be employed to solve the online learning problem efficiently.

Remark. Lemma 2 offers a guideline for analytically obtaining a closed-form gradient $\nabla r^{\psi}_{\theta}(z^*_t; W_t)$, even when a closed-form solution for z^*_t is unavailable. Utilizing Lemma 2 in CA-O Faster Learning is a win-win contribution from computational perspectives since it enhances precision and streamlines computation of the gradient. Moreover, CA-O Faster Learning accelerates computations, even in special cases where r^{ψ}_{θ} is concave. While certain cases require maximizing an intricate integral function over an ellipsoid set Θ_t, CA-O Faster Learning simplifies the process by only needing one evaluation $\nabla r^{\psi}_{\theta}(z^*_t; W_t)$, without relying on the concavity of r^{ψ}_{θ}. However, it is worth noting that the efficiency necessitates additional initial explorations to ensure optimal regret performance, which will be further discussed in Section 5.2.

5 Regret analysis

We are now ready to establish the regret bound for both CA-O Learning. and CA-O Faster Learning Algorithms. Define Regret of policy π as

$$\mathsf{Regret}_\pi(T) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} R_t(A^*_t; W_t, \theta^*) - R_t(A_t; W_t, \theta^*)\right] = \sum_{t=1}^{T} r_{\theta^*}(A^*_t; W_t) - r_{\theta^*}(A_t; W_t),$$

where A^*_t is the optimal store location action; that is, $A^*_t = A_t(z^*_t(\cdot;\theta^*))$, the action discretized from the optimal CA design based on true parameter θ^*. A_t is the action implemented at time t according to policy π. In our proposed algorithms, we take $A_t = A_t(z^*_t(\cdot;\theta_t))$. We measure the profit gap between A^*_t and A_t in regret, given that A^*_t represents the optimal discretized decision attainable by the retailer. In order to analyze the regret, we must quantify two gaps, as detailed in Lemma 3: The first, termed the *CA gap* and denoted by $\mathsf{Gap}_{\mathsf{CA}}$, is the disparity between the profit provided by the CA model and that from the discretized action. The second, referred to as the *learning gap*, is the difference between the profits implied by the same action under true and optimistic parameters.

Lemma 3. When θ^* is contained in the uncertainty set Θ_t, the regret contributed at time step t in our proposed algorithms can be decomposed as follows:

$$\begin{aligned} r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t) \le \underbrace{\big(r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t)\big)}_{\text{learning gap}} \\ - \underbrace{\big(r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t)\big)}_{\text{CA gap}} + \underbrace{\big(r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t) - r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t)\big)}_{\text{CA gap}}. \end{aligned} \tag{14}$$

In Section 5.1, we first quantify the CA gap, and show that the CA gap is moderate. Although Ouyang and Daganzo (2006) use several numerical examples of offline location problems to demonstrate that the CA gap is small, there lacks a universally valid theoretical upper bound for the CA gap. However, such a bound is particularly relevant in an online setting, where the CA gap may widen over time. We address this void in the literature by deriving an upper bound for the CA gap. Initially, we examine a single period, thereby omitting the time index t. Subsequently, in Section 5.2, we focus on the learning gap over the planning horizon, and then analyze the total regret.

5.1 CA Gap Analysis

In this subsection, we provide the intuition and technical assumptions to derive an upper bound of the CA gap. For brevity, we relegate detailed proof and a discretization procedure in Appendix C.1.

To determine the bounds for the CA gap, we are motivated by a pivotal alternative influence area function $z^s(x)$, constrained as a step function, such that $z^s(x) = |\mathcal{X}_j|,\ \forall x \in \mathcal{X}_j$. We refer to this alternative as the *step CA*. The definition of step CA allows us to decompose the CA gap for each influence area j into two parts: 1) the gap from the optimal CA $z^*(x)$ to the step CA, denoted by $\mathsf{Gap}_{j,o2s}$; 2) the gap from the step CA to the actual design obtained from the discretization procedure, denoted by $\mathsf{Gap}_{j,s2d}$. Thus, the CA gap can be bounded as $\mathsf{Gap}_{\mathsf{CA}} := \sum_j \mathsf{Gap}_{j,o2s} + \mathsf{Gap}_{j,s2d}$.

The approach to quantify $\mathsf{Gap}_{j,o2s}$ is to apply Taylor expansion at z^*, where the linear term vanishes due to the first-order condition. Referring to Cases 1–3, we assume that the profit density function $\psi(z)$ is twice differentiable and quasi-concave in z (which generally holds for mobile store location problems). Afterward,

$$\mathsf{Gap}_{j,o2s} := \int_{x\in\mathcal{X}_j} \big(\psi(z^*(x)) - \psi(z^s(x))\big)dx = -\int_{x\in\mathcal{X}_j} \frac{\psi''(\bar{z}(x))}{2\rho(x)}(z^s(x) - z^*(x))^2\rho(x)dx,$$

where $\bar{z}(x)$ is a convex combination of $z^s(x)$ and $z^*(x)$. In the above integral, $\frac{\psi''(z)}{\rho(x)}$ is the curvature of per-customer profit density function with a constant upper bound given by $\frac{|\psi''(z)|}{\rho(x)} \le \eta^{\psi}$. Since $z^s(x)$ can be regarded as the mean of $z^*(x)$ over $\mathcal{X}_j$, $\int_{x\in\mathcal{X}_j}(z^s(x) - z^*(x))^2\rho(x)dx$ measures the variability of $z^*(x)$ over $\mathcal{X}_j$. Combining the two terms together yields the following upper bound for $\mathsf{Gap}_{j,o2s}$:

Lemma 4. Define the variance of $z^*(x)$ over area $\mathcal{X}_j$ as $Var(z^*;\mathcal{X}_j) := \int_{x\in\mathcal{X}_j}(z^*(x) - \mathbb{E}[z^*])^2\rho(x)dx$. The profit gap between the optimal CA and the step CA for each influence area j is bounded as follows:

$$0 \le \mathsf{Gap}_{j,o2s} \le \frac{\eta^{\psi}}{2} Var(z^*;\mathcal{X}_j).$$

In general, if the underlying customer distribution profile $\rho(\cdot)$ is a slow-varying function (which is often the case in practice), so is the influence area function $z^*(\cdot)$. Then the variance of $z^*(\cdot)$ should be close to zero. We thus expect that $\mathsf{Gap}_{j,o2s}$ should be reasonably small.

We next quantify $\mathsf{Gap}_{j,s2d}$. Upon examining Cases 1–3, we summarize key profit function structures of mobile store location problems in Assumption 1.

Assumption 1 (Operations of Mobile Retail Stores)**.**

(I) Store locations $\boldsymbol{x} = \{x_1, x_2, ..., x_N\}$ are centroids of the influence areas.

(II) Both the revenue and the facility cost φ^f are affine functions of the daily sales at the store.

(III) The inventory replenishment cost φ^i is a concave function of both the truck routing distance and the daily sales at the store.

We adopt these assumptions given their relevance to the context of the problem, not solely for the sake of analytical convenience. Notably, Assumption 1(I) reflects real-world practices: stores are commonly located at influence area centroids because customers tend to prefer the nearest stores. Assumption 1(II) is widely used in facility location models. Assumption 1(III) is valid for Cases 2 and 3, although it is not entirely realistic for Case 1. In fact, the proof of Lemma 5 will show that our derived upper bound is always valid without Assumption 1(III) if we do not need to preserve the direction of the gap.

The magnitude of $\mathsf{Gap}_{j,s2d}$ depends on the variability in the profit function. Consider X as a random location in $\mathcal{X}_j$, and X' as a random location in $\mathcal{X}$. $Var(d(X', X); \mathcal{X}_j)$ denotes the variance of the random replenishment trip distance $d(X', X)$; $Var(\rho(X); \mathcal{X}_j)$ denotes the variance of the random demand density $\rho(X)$. With the additional value caps imposed in Assumption 2, Lemma 5 provides quantification of this gap.

Assumption 2 (Functional Boundedness)**.** The second derivative of the inventory replenishment cost φ^i with respect to inbound truck routing distance exists, and its absolute value is bounded from above by η^i. The second derivative of φ^i with respect to daily sales exists, and its absolute value is bounded from above by η^ρ.

Lemma 5. (I) Suppose Assumptions 1 and 2 hold, the profit gap between the step CA and the discrete design for each influence area j is bounded as follows:

$$0 \leq \mathsf{Gap}_{j,s2d} \leq \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j).$$

(II) Alternatively, relaxing Assumption 1(III) yields

$$|\mathsf{Gap}_{j,s2d}| \leq \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j)) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j).$$

Referring to the results of Lemma 5, one can expect that the gap between the step CA and the discrete design is likely to be mild, too. If the customer distribution $\rho(\cdot)$ is a slow-varying function, $Var(\rho(X); \mathcal{X}_j)$ should be close to zero, and the values of η^i and η^ρ will be small. In addition, if the influence areas are small (i.e., stores are densely deployed), $Var(d(X', X); \mathcal{X}_j)$ also tends to diminish. These conditions are commonly observed in practice.

Combining Lemma 4 and 5 immediately yields the following Theorem 1 of the CA gap. $\mathsf{Gap}_{\mathsf{CA}}$ is moderate since both $\mathsf{Gap}_{j,o2s}$ and $\mathsf{Gap}_{j,s2d}$ should be small. For ease of notation, we define the universal CA gap as

$$\beta_{\mathsf{CA}} := \sup_{\{\boldsymbol{\mathcal{X}}\}} \sum_j \left(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) + \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) \right),$$

where $\{\boldsymbol{\mathcal{X}}\}$ is the set of feasible partitions of influence areas divided by stores. β_{CA} represents the supremum obtained by considering all possible store locations and summing the CA gap for each influence area.

Theorem 1 (CA Gap)**.** Suppose Assumptions 1 and 2 hold. The CA gap incurred in each period is bounded as $0 \leq \mathsf{Gap}_{\mathsf{CA}} \leq \beta_{\mathsf{CA}}$. Furthermore, if relaxing Assumption 1(III), the CA gap is bounded as follows: $|\mathsf{Gap}_{\mathsf{CA}}| \leq \beta_{\mathsf{CA}}$.

Theorem 1 provides two bounds for the CA gap. The first bound guarantees that $\mathsf{Gap}_{\mathsf{CA}}$ is always non-negative, but holds under more restrictive assumptions. In what follows, we use the second bound in the regret analysis to account for a more general setting in mobile retail problem.

5.2 Regret analysis of learning algorithms

In this subsection, we prove the regret bounds for Algorithm CA-O Learning. and CA-O Faster Learning. Recall that the regret is decomposed into the *CA gap* and the *learning gap* in Lemma 3 where the learning gap is defined as $r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t)$. In this way, the gap between profits under different actions is transformed into the gap between the profit functions with parameters θ^* and θ_t under the same action $A(z_t^*(\cdot;\theta_t))$. Intuitively, when the difference between θ^* and θ_t is small, so is the regret incurred at time step t. Therefore, to derive the theoretical regret guarantees, we first aim to bound the uncertainty radius, which should be in the form of a norm of $\hat{\theta}_t - \theta^*$. With the assistance of the following two assumptions that are commonly acknowledged in the bandits literature, we are able to derive the bound for the radius of the uncertainty set.

Assumption 3 (Conditional Sub-Gaussianity)**.** There exists $\sigma > 0$ such that for every $t = 1, \ldots, T$ and for all $0 \le j \le N_t$ and $u \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(u\epsilon_{tj}) \mid \mathcal{H}_t] \le \exp(u^2\sigma^2/2)$.

Assumption 4 (Boundedness)**.** The following conditions hold:

(I) $r_{\max} := \sup_{A\in\mathcal{A},W\in\mathcal{W}} r_{\theta^*}(A;W) < \infty$ and $r_{\max} \ge 1$.
(II) $\hbar_f := \sup_{\theta\in\Theta,W\in\mathcal{W}} \|\nabla^2_\theta r^{\psi}_\theta(z^*(\cdot;\theta);W)\|_2 < \infty$.
(III) The maximal number of stores is $N_{\max} < \infty$.
(IV) $\beta_\Theta := \sup_{\theta\in\Theta} \|\theta\|_2 < \infty$.
(V) $\beta_\kappa := \sup_{x\in\mathcal{X},W\in\mathcal{W},A\in\mathcal{A}} \max\{\|\kappa(A,x;W(x))\|_2\} < \infty$.
(VI) φ^i is Lipschitz continuous on store daily sales (ρz) with modulus α_i.
(VII) $\alpha_f := \sup \left|\frac{\partial\varphi^f}{\partial(\rho z)}\right| < \infty$.

Essentially, the σ-subgaussianity in Assumption 3 regulates that the tails of the response noise ϵ_t decay at reasonably fast rate. Assumption 4 offers bounding constants that are instrumental in the derivation of regret guarantees. We now proceed to give a high-level idea of constructing the uncertainty set for parameter θ^*. Recall that g_{sj} denotes the gradient $\nabla f_\theta(A_s;W_s,\mathcal{X}_{sj})$ and λ is used in the squared-loss function ℓ. Define $\xi_t = \sum_{s=1}^{t-1}\sum_{j=1}^{N_s} \epsilon_{sj} g_{sj}$. We obtain the following lemma by reformulating (Oracle) into $\|\hat{\theta}_t - \theta^*\|_{V_t} = \|\xi_t - \lambda\theta^*\|_{V_t^{-1}}$ and applying Cauchy-Schwarz inequality, with detailed proof in Appendix C.2.

Lemma 6. It is established that $\|\hat{\theta}_t - \theta^*\|_{V_t} = \|\xi_t - \lambda\theta^*\|_{V_t^{-1}} \le \|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta$.

In light of Lemma 6, $\|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta$ provides an upper bound for a proper choice of the radius of the uncertainty set, i.e., $\|\hat{\theta}_t - \theta^*\|_{V_t}$. This inequality motivates us to provide a high-probability upper bound for the stochastic term $\|\xi_t\|^2_{V_t^{-1}}$ using the definition of σ-subgaussian noise in Assumption 3, as demonstrated in the following lemma. The high-level idea is to construct a non-negative supermartingale $M_t(x) = \exp(\langle x,\xi_t\rangle - \frac{\sigma^2}{2}\|x\|^2_{V_t-\lambda I})$ and then apply method of mixtures, the proof of which is in Appendix C.2.

Lemma 7. For any $\delta \in (0,1]$, we have

$$\mathbb{P}\left(\exists t \ge 1, \|\xi_t\|^2_{V_t^{-1}} \ge \sigma^2\left(2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)\right)\right) \le \delta.$$

Lemma 7 shows that $\|\xi_t\|^2_{V_t^{-1}}$ is in order $O(\log(t))$ with probability at least $1-\delta$. Combining Lemma 6 and 7 provides an upper bound for $\|\hat{\theta}_t - \theta^*\|_{V_t}$. Further, by establishing an upper bound

for $\det(V_t)/\lambda^d$ with Assumption 4(V), we immediately obtain the following result on the radius of the parameter uncertainty set:

Lemma 8 (Radius of uncertainty set)**.** Assuming that Assumptions 3 and 4 are in force, it holds with probability at least $1-\delta$ that, for all $t \in [T]$,

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t,$$

where $\gamma_t = \sqrt{\lambda}\beta_\Theta + \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{(t-1)|\mathcal{X}|^2\beta_\kappa^2}{\lambda d}\right)}$.

Lemma 8 shows that the radius of the uncertainty set $\|\hat{\theta}_t - \theta^*\|_{V_t}^2$ is in order $O(\log(t))$. At each time step t, Algorithm CA-O Learning. and CA-O Faster Learning solve the optimization problem over the confidence set $\|\hat{\theta}_t - \theta\|_{V_t} \leq \gamma_t$. According to Lemma 8, θ^* falls into this confidence set with high probability, which also implies that the algorithm finds an optimistic solution.

According to the CA gap result in Theorem 1, for all $1 \leq t \leq T$, we obtain

$$|r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t)| \leq \beta_{\mathsf{CA}}, \quad |r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t)| \leq \beta_{\mathsf{CA}}.$$

Define event $\mathcal{E}_t = \{\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t\}$. When $\mathcal{E}_t$ holds, based on the regret decomposition (14), the regret accumulated at time step t can be bounded as follows:

$$\begin{aligned}&(r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t)\\ \leq&(r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t) + 2\beta_{\mathsf{CA}}.\end{aligned}$$

It remains to bound the learning gap, represented by the first term on the right-hand side of the inequality. While our observation is on the demand, the learning gap is measured in terms of profit. To address this discrepancy, the following lemma shows that a profit gap can be bounded by a demand gap.

Lemma 9. Suppose Assumption 4 holds. There exist constants $L_r > 0$ such that for every $\theta \in \Theta$, and for every subzone $\mathcal{X}_j \subseteq \mathcal{X}$, $A \in \mathcal{A}$, $W \in \mathcal{W}$, and $1 \leq t \leq T$,

$$|r_\theta(A;W,\mathcal{X}_j) - r_{\theta^*}(A;W,\mathcal{X}_j)| \leq L_r|f_\theta(A;W,\mathcal{X}_j) - f_{\theta^*}(A;W,\mathcal{X}_j)|,$$

where $L_r = \bar{r} + |\mathcal{X}|\alpha_i + |\mathcal{X}|\alpha_f$.

Lemma 9 implies that the learning gap is bounded by $L_r\sum_{j=1}^{N_t}|(\theta_t - \theta^*)^\top g_{tj}|$. Combining Lemmas 6–9, the following theorem establishes the regret bound for Algorithm CA-O Learning.. In the regret analysis, we first show that the algorithm finds an optimistic solution at each step, and then quantify the profit gap incurred by both the learning gap and the CA gap. The detailed proof is provided in Appendix C.2.

Theorem 2 (Regret of Algorithm CA-O Learning.)**.** Assume Assumptions 1–4 are in force. Let $\delta \in (0,1)$. With probability at least $1-\delta$, we can bound $\mathsf{Regret}_\pi(T)$ as follows:

$$\mathsf{Regret}_\pi(T) \leq r_{\max}t_0 + 2r_{\max}L_r\gamma_T\sqrt{2N_{\max}dT\log\left(\frac{d\lambda + T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\beta_{\mathsf{CA}}T.$$

Next we characterize the regret bound for Algorithm CA-O Faster Learning. To expedite the learning process, we utilize the first-order approximation, which incurs additional error of $O(\|\theta^* - \hat{\theta}_t\|_2^2 + \|\hat{\theta}_t - \theta_t\|_2^2)$. This approximation performs better when the estimated parameter gets closer to the true parameter. Thus, we first explore t_0^F time periods to guarantee that the estimated parameter is close to the true parameter, then we simultaneously explore and exploit using the faster learning algorithm. We impose the following assumption on the context divergence, which makes it possible to gather information from the exploration periods.

Assumption 5 (Context Diversity)**.** During the exploration periods, the retailer adopts a randomized policy π to generate the number of stores $N(A)$ and location decisions A from uniform distributions. There exists a constant $0 < \underline{\lambda} < \infty$ such that the eigenvalues can be bounded by

$$\lambda\left(\mathbb{E}_{A\sim\pi(\mathcal{A}),W}\left[\sum_{j=1}^{N(A)}\int_{\mathcal{X}_j}\kappa(A,x;W(x))dx\int_{\mathcal{X}_j}\kappa(A,x;W(x))^\top dx\right]\right)\geq\underline{\lambda}.$$

The two algorithms require different numbers of exploration periods. Algorithm CA-O Learning. needs only $O(1)$ initial explorations, while Algorithm CA-O Faster Learning requires more to ensure a good approximation. We define $t_0^F = \max\left\{\frac{\sqrt{T}}{(1-\delta)\underline{\lambda}}, \frac{(\log(d)-\log(\delta))|\mathcal{X}|^2\beta_\kappa^2}{(\delta+(1-\delta)\log(1-\delta))\underline{\lambda}}\right\}$. Under Assumption 5, Lemma 12 in Appendix C.2 demonstrates that, after t_0^F exploration periods, the distance between $\hat{\theta}$ and θ^* is at most $O(\frac{1}{\sqrt{T}})$ with high probability. With this order of approximation, we quantify the regret of Algorithm CA-O Faster Learning in Theorem 3.

Theorem 3 (Regret of Algorithm CA-O Faster Learning)**.** Assume Assumptions 1–5 are in force. Let $\delta \in (0,1)$. With probability at least $1-2\delta$, we can bound $\mathsf{Regret}_\pi(T)$ as follows:

$$\mathsf{Regret}_\pi(T) \leq r_{\max}t_0^F + 2r_{\max}L_r\gamma_T\sqrt{2N_{\max}dT\log\left(\frac{d\lambda+T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\hbar_f\gamma_T^2\sqrt{T} + 2\beta_{\mathsf{CA}}T.$$

Theorems 2 and 3 suggest that the regret decomposes primarily into $O(d\sqrt{T})$ and $O(\beta_{\mathsf{CA}}T)$. The former arises from inherent parameter uncertainty in the learning algorithm, while the latter results from using the CA approach to address computational and analytical challenges. However, our numerical experiments reveal that the regret remains sublinear. Furthermore, when our estimator is close to the true parameters, the two CA gaps in (14) largely cancel each other out, making the impact of the CA gap minimal.

We conclude this section on regret analysis by highlighting the contributions of our proposed algorithms. First, our algorithms resolve the complexity raised by the task of demand learning in location models. We adopt the CA approach to convert combinatorial action and objective functions into continuous models and thus simplify the action space. Second, we address the computational challenges in optimistic algorithms due to the potential nonconvexity. The continuous functions provided by CA enable the application of first-order approximation, which, coupled with the closed-form gradient expression, jointly enhances the efficiency of Algorithm CA-O Faster Learning, even if the optimization is non-convex or lacks a closed-form expression of $z^*(\cdot)$. Meanwhile, this algorithm maintains the same order of regret with respect to T, indicating minimal precision loss. Moreover, beyond the realm of mobile retail store locations, the efficiency of CA-O Faster Learning holds potential for other bandits problems with intricate structures, such as online vehicle routing.

6 Computational studies and managerial insights

To evaluate the empirical efficacy of the algorithms proposed in this paper, we conduct two main numerical experiments. Section 6.1 presents a comparison using synthetic data in a mobile store location problem to highlight the benefits of approximation. In Section 6.2, we conduct a case study based on real-world data from Toronto. Section 6.3 then builds on these findings, contrasting our algorithms' performance against established benchmarks to underscore the value of mobility and online learning in urban retail contexts.

6.1 Efficacy of the Faster Learning algorithm

As discussed in Case 1, both CA-O Learning. and CA-O Faster Learning can be applied to a mobile store location problem when a closed-form solution exists. In this subsection, we specifically compare these two algorithms from two aspects - the regret performance and the computational time.

We set up numerical experiments as follows. We set $\bar{r} = 6$, $a^f = 2$, $b^f = 100$, $\bar{c}_t = 0.03$, $S = 50$. The contextual covariates $W(x) \in \mathbb{R}^{50}$ are generated by Gaussian kernel functions. Customer demands are distributed in a square region $\mathcal{X}_t := [0, 1] \times [0, 1]$. The observational noise is drawn from a Normal distribution with mean zero and standard deviation being 50% of the actual demand of each influence area. Both algorithms run for 4000 times using an AMD EPYC 7532 processor at 2.4GHz.

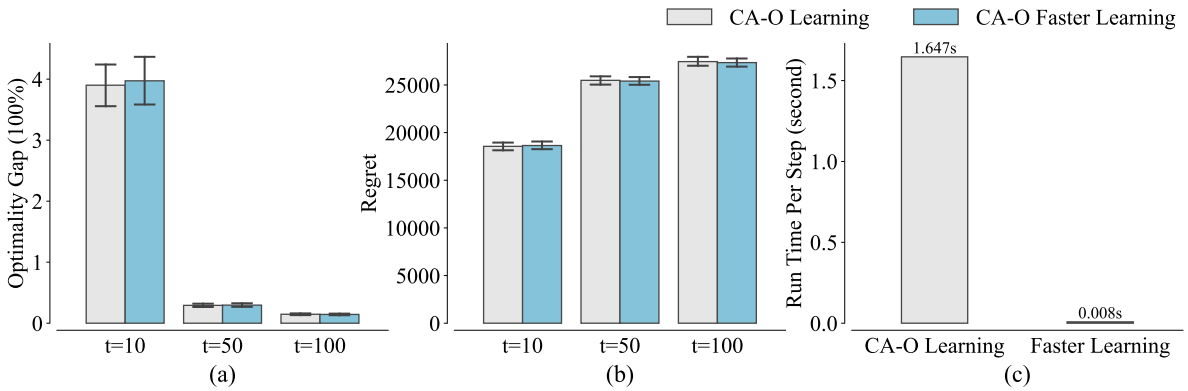


Figure 3: The comparison of CA-O Learning. and CA-O Faster Learning: (a) optimality gap; (b) regret; (c) run time per round. (a) and (b) are evaluated at three different rounds. (c) is evaluated by the average value of the planning horizon T.

We compare the two algorithms using three metrics as the bar-charts in Figure 3 illustrate. Specifically, the first metric is the optimality gap, defined as

$$\text{optimality gap} := \frac{r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t)}{r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t)} \times 100\%.$$

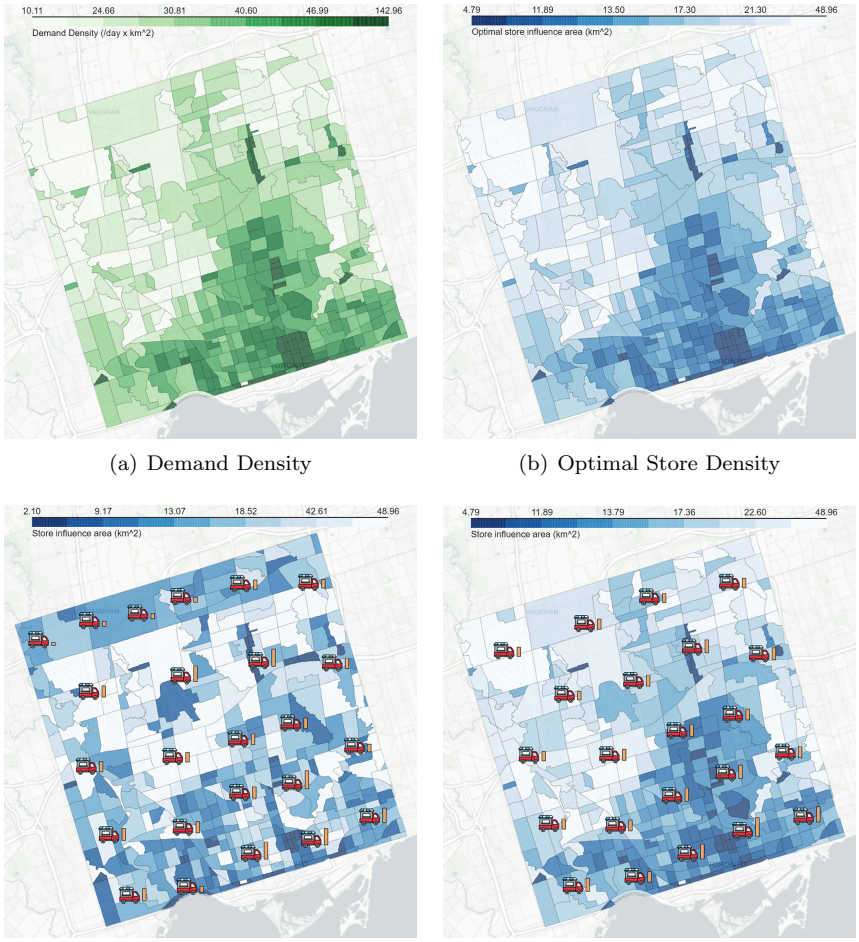
In other words, the optimality gap is the relative gap between the optimal reward obtained when knowing the true parameter θ^* and the reward obtained based on the optimistic estimator θ_t. Figure 3(a) shows that both algorithms learn very fast: The optimality gaps are under 5% after 10 rounds and under 1% after 100 rounds. Furthermore, Figures 3(b) and (c) show that CA-O Faster Learning is able to achieve similar regrets as CA-O Learning. at a 95% confidence level, but boasts about 200 times higher computational efficiency.

6.2 Algorithmic advantages in real-world applications

To illustrate how Algorithm CA-O Faster Learning can be applied to solve a real-world problem without a closed-form $\psi_\theta(z_t^*(x;\theta); W_t(x))$, we examine the mobile retail problem specified in Case 3. Following the CA model in (13), note that we are not able to obtain a closed-form maximizer. The optimization involved in CA-O Learning. would be intractable. However, Algorithm CA-O Faster Learning can be used to overcome this computational hurdle.

The experiment setting is as follows:. We set $\bar{r} = 6$, $a^f = 2$, $b^f = 400$, $c^t = 3$, $S = 50$, $c_0 = 0.5$. The context information of 23-dimensional data is obtained from real-world data in an urban area of Toronto, Canada. In this context, 10 spatial attributes are selected from 2021 Census of Population (`statcan.gc.ca`). The other 13 attributes are time series data for the year 2022, including temperature, precipitation, speed of gust, Fisher commodity price index for grocery and energy, historical retail trade sales, and 7 indicators for weekdays. Figure 4(a) visualizes the demand density. The observational noise is drawn from a normal distribution with a mean of zero and a standard deviation of 50% of the actual demand in each influence area.

The decision-maker improves the store layout in the process of online learning. The update frequency of store locations is set daily. We set $t_0^F = 1$ for an initial exploration. After day $t = 2$, when the online learning algorithm just starts, the store layout provided by our algorithm in Figure 4(c)



(a) Demand Density (b) Optimal Store Density

(c) Store Locations on Day $t = 2$ (d) Store Locations on Day $t = 30$

Figure 4: Spatial visualization of ground truth information and stores locations in Toronto case study. The deeper green color in (a) indicates higher demand density. The deeper blue color in (b)(c)(d) indicates smaller store influence area (i.e., higher stores density). The length of the vertical orange bar next to mobile store indicates the daily sales of the store.

deviates significantly from the ground truth optimal store density shown in Figure 4(b). Nevertheless, by day $t = 30$ when we have a better estimation of the demand, the store layout decided by our algorithm is already near-optimal, as shown in Figure 4(d). At this point, the profit is quite close to the maximal profit. The figures show that a significant profit increment is achieved within only a few rounds. This result demonstrates that CA-O Faster Learning quickly learns and converges.

To show the regret performance, we run Algorithm CA-O Faster Learning 200 times. Figure 5(a) shows a clear sublinear trend of the regret. This numerical performance is better than our expectation from Theorem 3, in which a linear term $2\beta_{\mathsf{CA}}T$ shows up to bound the CA gap. This favorable numerical result is consistent with our preceding reasoning in Section 5.2 that the actual effect of CA gaps is minimal. Indeed, Figure 5(b) provides visualization of how the value of the two CA gaps incurred by the optimistic solution and by the optimal solution are of similar magnitude, and of how these two gaps largely offset each other because they arise in the opposite direction on the basis of (14). Such offset is why the linear term in the theoretical analysis vanishes in our experiment, and the error resulting from the CA gap is much smaller than the conservative bound given in Theorem 3.

We also are interested in the performance comparison between CA-O Faster Learning and the baseline algorithms. In particular, we use as baselines a class of online learning algorithm named explore-

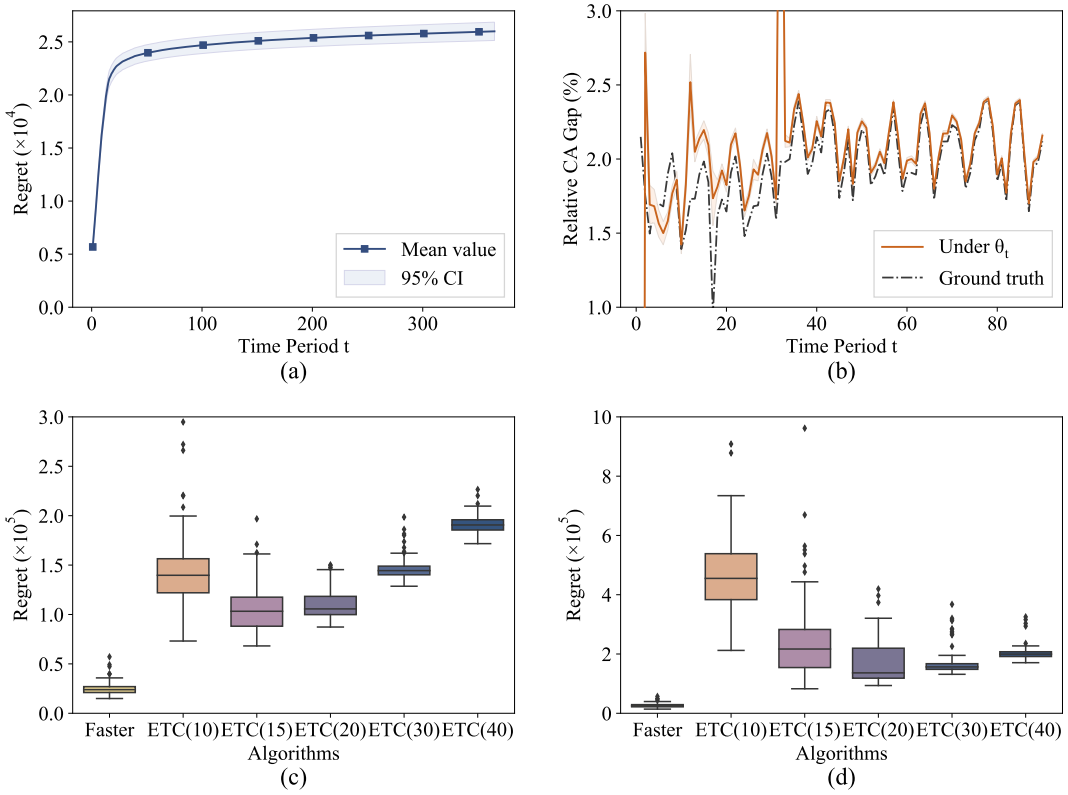


Figure 5: Regret performance of Algorithm CA-O Faster Learning. (a) Mean value and 95% confidence interval of regret; (b) Two CA gaps: during online learning and under the ground truth; (c) Regret of algorithms at day $t = 90$; (d) Regret of algorithms at day $t = 365$. "Faster": CA-O Faster Learning; "ETC(t_0')": Explore-then-commit algorithm, with different exploration periods t_0'.

then-commit (ETC) (Lattimore and Szepesvári 2020, Chapter 6). An ETC algorithm first explores by randomly designing a facility layout within a fixed number of rounds t_0 and then exploits by committing to the θ estimated during exploration. We test the ETC algorithms with various t_0 values ($t_0 = 1, 2, 4, 6, 8$) and compare the regrets of the ETC algorithms with the regret of the CA-O Faster Learning Algorithm. The superiority of CA-O Faster Learning is clearly shown in Figure 5(c)(d), where the regrets are accumulated both over the first quarter (including the first t_0 time periods) and over the whole year. In the latter case, the optimal exploration period for ETC is $t_0 = 20$ with a regret of 1.726×10^5. In contrast, CA-O Faster Learning results in a regret that is smaller than the regret from the best ETC algorithm by 67.5% on the day $t = 365$.

6.3 The value of learning & mobility

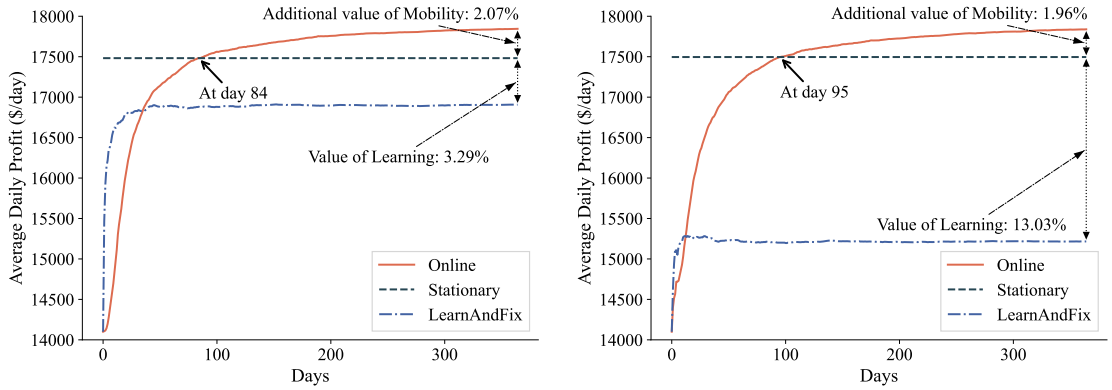
The mobile retail store business model offers advantages in two main aspects: demand learning and store mobility. To quantify these benefits, we extend experiments in Section 6.2. Additionally, we assess the impact of varying observation noise.

Value of Learning. The value of learning arises from resolving demand uncertainty. To distinguish this from the benefits of store mobility, we maintain fixed locations when assessing the value of learning. We consider two benchmark retail models:

i. "Stationary Retail", where store locations, determined at the beginning of the planning period based on the yearly average of ground truth demand density data, remain unchanged.

ii. "Learn and Fix", which involves a one-day demand exploration followed by fixed store location decisions for the remainder. Given only one exploration step, the estimation is inevitably imprecise.

We employ the average daily profit from the start to day t as our evaluation metric for retail models. With minimal demand uncertainty, Figure 6(a) reveals a profit gap between the "Stationary Retail" and "Learn and Fix" of 3.29% by the end of the year, highlighting the benefits of demand learning. Conversely, under high demand uncertainty as in Figure 6(b), the gap expands to 13.03%. This gap expansion is expected, because resolving larger uncertainties yields greater profit. Although our learning algorithms are not directly applied here, these gaps illustrates the importance of demand learning in the mobile store location problem. Notably, mobile retail stores offer superior demand learning capabilities compared to traditional ones where assuming ground truth information may be unrealistic.



(a) Smaller Observational Noise (b) Larger Observational Noise

Figure 6: The value of demand learning and store mobility. "Smaller" and "larger" noise indicate the noise with a standard deviation of 20% and 80% of the actual demand, respectively.

Additional Value of Mobility. While "Stationary Retail" provides optimal parameter estimation, it lacks adaptability to changing contexts. In contrast, the Algorithm CA-O Faster Learning leverages mobility to adapt to dynamic contexts and successively refines parameter estimation. As shown in Figure 6(a), the additional profit generated by Algorithm CA-O Faster Learning compared to "Stationary Retail" converges to 2.07%, representing the added value of mobility. This surplus is almost consistent at 1.96% in Figure 6(b). It's important to note that in real-world scenarios, this value could be much greater. The "Stationary retail" model, which assumes perfect demand knowledge, is idealistic, while Algorithm CA-O Faster Learning offers a practical approach.

Impact of Noise in Observation. Each of the three business models is tested under observational noise affecting demand, with a standard deviation of 20% of the actual demand in Figure 6(a) and 80% in Figure 6(b). In the scenario with the higher noise level, which is typical in practice, there is a significant rise in the value of learning. In Figure 6(a), the curves of mobile stores and "Stationary Retail" intersect by day 84. Conversely, Figure 6(b) shows an intersection by day 95. Although it takes more days for Algorithm CA-O Faster Learning to surpass "Stationary Retail" in cumulative profits compared to the scenario with the smaller noise level (Figure 6(a)), the overall profits decrease by a mere 0.11%. These results highlight the advantages of mobile retail stores and the robustness of our algorithms against noise.

7 Conclusion

This paper is motivated by the growing trend of mobility in urban retail. The emergence of mobile retail stores suggests that the store location design can be made at the operational level with short-time adjustments, inspiring us to expand the scope of facility location problems into an online setting. We develop an online learning framework for OFL problems with contextual information. The retailer designs mobile store locations while learning from past observations. To overcome the challenge stemming from the infinite-dimensional action space and the dependence between actions and observations, we propose CA-O Learning. Algorithm by combining the CA technique with an optimistic optimization problem. Moreover, we propose CA-O Faster Learning to handle a more general class of model structures and significantly improve the computational efficiency. The theoretical regret characterization reveals that both algorithms guarantee low regret. Through our experiments, we verify this low regret for both algorithms and also highlight the high computational efficiency of CA-O Faster Learning. Moreover, our case study underscores the significant benefits of demand learning and the inherent mobility of mobile retail stores when using CA-O Faster Learning.

While our research centers on mobile retail store locations, the algorithms developed have applicability extending beyond this. The transition from stationary locations to enhanced mobility in urban dynamics, as evidenced by the examples below, further motivates and justifies this paper:

- *Mobile chargers.* The electric vehicle (EV) industry faces challenges meeting the increasing charging demand. Traditional stationary infrastructure, burdened by long construction times and high capital investment, struggles to adapt to the evolving landscape of EV technologies and spatial distribution of new EV owners. These concerns have motivated the adoption of mobile charging in the form of battery-equipped robots or vans. Such innovations, advocated by industry leaders such as Volkswagen (IDTechEx 2020), offer rapid deployment and adaptability at reduced costs.
- *Micro-depots.* DPD Germany launched a new form of city logistics in Dresden by placing containers in parking areas as local micro-depots (DPD 2023). These micro-depots serve as a storage and transshipment point, facilitating the outbound last-mile deliveries by cargo bikes or crowdsourced mobility. Their ease of deployment, combined with lower emissions and better access to narrow city streets compared to conventional vans, highlights the potential of micro-depots.

Our work represents an early attempt to deploy mobile stores in an online fashion and we consider this paper a prompt for an open thread. Several potential extensions to our work are worth investigating. First, due to the limited capacity of mobile stores, it becomes important to decide what products to display and how many to stock based on consumer behavior. As consumers may become loyal to these stores over time, understanding and incorporating their preferences in location design becomes increasingly important. Integrating consumer choice behavior into the location design problem, especially when considering the long-term effects of store deployment, is a substantial challenge. Furthermore, when consumer preferences towards different products are unknown, how to dynamically adjust the assortment while simultaneously learning about these preferences is an intriguing area for exploration. Second, many other important and practical business constraints also need to be considered. For example, moving the location of mobile facilities may incur a moving or routing cost, which is not considered in our mobile store location problem. Future work on these extensions will broaden and deepen our understanding about mobile facility deployment toward a vibrant urban future energized by data-driven and agile services.

A Notation

Table 1: Notation.

Symbol	Description
t	round (or time) in the online learning process
T	number of rounds in total
$\mathcal{X}, \mathcal{X}_t$	entire region at time t
$\mathcal{X}_{tj}$	influence area of store j at time t
x	a location in the region
$\boldsymbol{x}_t$	a set of store locations at time t
x_{tj}	store location of influence area j at time t
N_t	number of stores for time t
$\mathcal{W}_t$	contextual information set for time t
$W_t(x)$	local contextual covariates at location x at time t
$\mathcal{A}_t$	action set for time t
$A_t, A_t(\boldsymbol{x}_t, \boldsymbol{\mathcal{X}}_t)$	discrete store location decisions at time t
Y_t	observed demand served by the store at location x_{tj} at time t, i.e., response vector
R_t	profit received at time t
θ_t	estimation of parameter vector at time t
θ^*	ground truth parameter vector
Θ_t	uncertainty set for parameter θ at time t
$r_\theta(A_t; W_t)$	expected profit at round t given action A_t
$r_{\theta^*}(A_t; W_t)$	expected true profit at round t given action A_t
$\rho(x), \rho_\theta(A_t, x; W_t(x))$	demand density for x given action A_t
$\kappa(A_t, W_t(x))$	kernel vector of the features at location x given action A_t and context $W_t(x)$
$\varphi_\theta(A_t, x; W_t)$	expected profit density around location x given action A_t at time t
$\bar{r}$	average revenue per customer
φ^i, φ^f	inventory replenishment cost and facility cost density function, respectively
$d(\cdot, \cdot)$	distance function
$f_\theta(A_t; W_t, \mathcal{X}_{tj}), f_{sj}(\theta)$	fitted demand given parameter θ and action A_t at time t
ϵ_{tj}	observational noise of demand served by the store located at x_{tj} at time t
$\mathcal{Z}_t$	a set of non-negative and continuous functions for influence area function
$z_t(x), z_t(x; \theta)$	influence area function for location x at time t, the decision of CA model
$z^*, z_t^*(\cdot; \theta)$	CA recipe, the optimal solution of CA under parameter θ
$\psi_\theta(z_t(x); W_t(x))$	continuous profit density provided by CA approach at location x at time t given action z_t
S	the volume of each refill for inventory replenishment
c_t	the routing cost of trucks in inventory replenishment per kilometer of travel
β_{TSP}	traveling salesman problem (TSP) constant
a^f	goods-handling cost per item of mobile retail stores
b^f	fixed opening cost per store
$A(z^*(\cdot; \theta))$	discrete store location decisions tranlated from CA recipe $z^*(\cdot, \theta)$
$A_t^*, A_t(z_t^*(\cdot; \theta^*))$	discrete store location decisions from CA recipe based on true parameter θ^*
$r_\theta^\psi(z; W)$	approximate profit function from the CA model given action z
ℓ	squared-loss function
λ	l_2-norm regularization parameter for the loss function
g_{sj}	gradient of fitted demand $f_{sj}(\theta)$ with respect to θ
V_t	the design matrix at time t to construct ellipsoid uncertainty set
γ_t	radius of ellipsoid uncertainty set
t_0, t_0^F	number of rounds to randomly explore actions in Algorithm CA-O Learning. and CA-O Faster Learning, respectively
β_{CA}	a universal upper bound for CA gap
$\mathcal{E}_t$	the event that θ^* is contained in the uncertainty set Θ_t
L_r	a Lipschitz constant for profit function r_θ with respect to demand

B Proofs in Section 4

Proof of Lemma 1. Let $\nu = \nabla r_{\hat{\theta}_t}^\psi(z_t^*(\cdot; \hat{\theta}_t); W_t)$ and $u = \theta - \hat{\theta}_t$, the optimization problem is

$$\max_{u: \|u\|_{V_t}^2 \le \gamma_t^2} \nu^\top u.$$

The Lagrangian function is $\mathcal{L}(u,\eta) = \nu^\top u - \eta(\|u\|_{V_t}^2 - \gamma_t^2)$ where $\eta \geq 0$. Then the gradient of $\mathcal{L}(u,\eta)$ equals

$$\nabla\mathcal{L}(u,\eta) = \nu - 2\eta V_t u = 0,$$

which gives $u = \frac{1}{2\eta}V_t^{-1}\nu$. Complementary slackness gives $\|u\|_{V_t}^2 = \gamma_t^2$, which implies that $\eta = \frac{\|\nu\|_{V_t^{-1}}}{2\gamma_t}$. Therefore, we have $u = \frac{1}{2\eta}V_t^{-1}\nu = \frac{\gamma_t V_t^{-1}\nu}{\|\nu\|_{V_t^{-1}}}$. The optimal θ_t has a closed-form solution that

$$\theta_t = \hat{\theta}_t + \gamma_t \frac{V_t^{-1}\nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot;\hat{\theta}_t);W_t)}{\|\nabla r^{\psi}_{\hat{\theta}_t}(z_t^*(\cdot;\hat{\theta}_t);W_t)\|_{V_t^{-1}}}.$$

□

Proof of Lemma 2. Consider the general expression of the continuous profit density function (6):

$$\psi_\theta(z(x);W(x)) = \bar{r}\rho(x) - \varphi^i\Big(\rho(x)z(x), z(x); W(x)\Big) - \varphi^f\Big(\rho(x)z(x), z(x); W(x)\Big),$$

where $\rho(x)$ is an abbreviation of $\rho_\theta(z(x), x; W(x))$, depending on $z(x)$, $W(x)$ and θ. Note that $\psi_\theta(z(x);W(x))$ is a function of $\rho(x)$, $z(x)$, $W(x)$ and θ. Since $\psi_\theta(z(x);W(x))$ is differentiable on $z(x)$, one can easily obtain optimal solution $z^*(x;\theta)$ by solving the first order condition

$$\frac{\partial\psi_\theta}{\partial z} = \frac{\partial\psi_\theta}{\partial(\rho(x))}\frac{\partial\rho(x)}{\partial(z^*(x;\theta))} + \frac{\partial\psi_\theta}{\partial(z^*(x;\theta))} = 0, \tag{15}$$

where $\frac{\partial\psi_\theta}{\partial(\rho(x))}$ means the partial derivative of $\psi_\theta(z(x);W(x))$ with respect to the argument $\rho(x)$, and the solution $z^*(x;\theta)$ is a function of θ at each point x. So far the maximal profit ψ_θ^* is obtained by

$$\psi_\theta^* = \psi_\theta\left(z^*(x;\theta);W(x)\right).$$

By the chain rule, the gradient of ψ_θ^* with respect to θ is represented as

$$\begin{aligned}\nabla_\theta\psi_\theta^* &= \frac{\partial\psi_\theta}{\partial(\rho(x))}\left(\nabla_\theta\rho(x) + \frac{\partial\rho(x)}{\partial(z^*(x;\theta))}\nabla_\theta z^*(x;\theta)\right) + \frac{\partial\psi_\theta}{\partial(z^*(x;\theta))}\nabla_\theta z^*(x;\theta)\\ &= \frac{\partial\psi_\theta}{\partial(\rho(x))}\nabla_\theta\rho(x) + \left(\frac{\partial\psi_\theta}{\partial(\rho(x))}\frac{\partial\rho(x)}{\partial(z^*(x;\theta))} + \frac{\partial\psi_\theta}{\partial(z^*(x;\theta))}\right)\nabla_\theta z^*(x;\theta).\end{aligned} \tag{16}$$

At the right-hand side of (16), the second term vanishes because $z^*(x;\theta)$ satisfies (15). Thus, we have

$$\nabla_\theta\psi_\theta^* = \frac{\partial\psi_\theta}{\partial(\rho(x))}\nabla_\theta\rho(x).$$

Note that $\rho_\theta(z^*(x;\theta), x; W(x)) = \theta^\top\kappa(z^*(x;\theta), W(x))$ implies $\nabla_\theta\rho(x) = \kappa(z^*(x;\theta), W(x))$, and the partial derivative of the profit density function $\psi_\theta(z(x);W(x))$ with respect to $\rho(x)$ is as follows:

$$\frac{\partial\psi_\theta}{\partial(\rho(x))} = \bar{r} - z(x)\left(\frac{\partial\varphi^i}{\partial(\rho(x)z(x))} + \frac{\partial\varphi^f}{\partial(\rho(x)z(x))}\right).$$

Therefore, the gradient of the maximal profit ψ_θ^* is provided by

$$\nabla_\theta\psi_\theta^* = \left[\bar{r} - z^*(x;\theta)\left(\frac{\partial\varphi^i}{\partial(\rho(x)z^*(x;\theta))} + \frac{\partial\varphi^f}{\partial(\rho(x)z^*(x;\theta))}\right)\right]\kappa(z^*(x;\theta), W(x)).$$

It concludes that one only needs the value of $z^*(x;\theta)$ rather than any derivative of $z^*(x;\theta)$. Recall that $z^*(x;\theta)$ can be evaluated pointwise through the first-order condition, facilitated by the convenience of CA. Therefore, the gradient is easy to compute even in the absence of a closed-form solution $z^*(x;\theta)$, meaning that $z^*(x;\theta)$ is an implicit function but always numerically computable. □

C Proofs in Section 5

Proof of Lemma 3. When θ^* is contained in the uncertainty set Θ_t, we have

$$
\begin{aligned}
& r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t) \\
=& r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) + r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t) \\
\le& r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) + r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t) \\
=& r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) + r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t) \\
& + r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) \\
=& \underbrace{\left(r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t)\right)}_{\text{learning gap}} \\
& - \underbrace{\left(r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t)\right)}_{\text{CA gap}} + \underbrace{\left(r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t) - r_{\theta_t}(A(z_t^*(\cdot;\theta_t));W_t)\right)}_{\text{CA gap}},
\end{aligned}
$$

where the inequality holds due to θ_t being an optimistic CA solution, ensuring that $r^{\psi}_{\theta^*}(z_t^*(\cdot;\theta^*);W_t) \le r^{\psi}_{\theta_t}(z_t^*(\cdot;\theta_t);W_t)$ when $\theta^* \in \Theta_t$. □

C.1 Proofs in Section 5.1

To prove Theorem 1, we first introduce the necessary preliminaries to analyze the gap. Next, we analyze the gap between the optimal CA and the step CA, the gap between the step CA and the discrete design. Finally, we conclude by the proof of Theorem 1.

Preliminaries of the CA Gap.

The continuous influence area design $z_t^*(x)$ can be discretized for implementation of single-period store locations (See Ouyang and Daganzo (2006) for a discretization procedure, which is beyond the scope of our paper). We also summarize the procedure in the following paragraph for readers' convenience.

The objective of discretization is to determine discrete store locations such that the influence areas of stores closely align with the values provided by the continuous function $z(x)$ across the entire space $\mathcal{X}$. The discretization procedure, as outlined in the referenced paper, can be summarized as three key steps: To translate a CA recipe $z(x)$ into N discrete influence areas, where $N \approx \int_{\mathcal{X}}[z(x)]^{-1}dx$, we represent these influence areas as circular disks centered at N arbitrary locations in the initial step. The size of each disk is determined by the value of $z(x)$ at its center. The second step is to iteratively *slide* and *shrink* the N disks to eliminate overlap. In each iteration, disks are slid due to repulsive forces from other overlapping disks to prevent overlap, and from boundary forces if a disk is outside of the region $\mathcal{X}$. At the same time, disk sizes are adjusted according to the value of $z(x)$ at each disk's center. The shrinking is done by simultaneously shortening the radii for disks each centering at a location x. Iterations end once the non-overlapping disks collectively cover most of $\mathcal{X}$ without extending beyond it, with each disk k centered at location x_k. In the third step, we partition $\mathcal{X}$ into N influence areas using a weighted-Voronoi tessellation. Specifically, each small patch of space is allocated to an influence area $\mathcal{X}_j$ with the rule $j = \arg\min_k\{\|x - x_k\|/\sqrt{z(x_k)}\}$, where x is the center of the patch. Consequently, we achieve a discrete design where the influence areas fully span the service space $\mathcal{X}$, with each area containing one disk and a store located at its center. This design is notably near-optimal, as discovering the globally optimal design of locations in the continuous domain is generally infeasible.

Our examination of the bounds for the CA gap draws inspiration from the proof of the CA gap direction provided by Ouyang and Daganzo (2006). In particular, a pivotal concept in their proof involves introducing the alternative influence area function $z^s(x)$, which we refer to as the *step CA*. As

aforementioned, the CA gap can be decomposed as $\mathsf{Gap}_{\mathsf{CA}} := \sum_j \mathsf{Gap}_{j,o2s} + \mathsf{Gap}_{j,s2d}$. The subsequent two parts present bounds for these two gaps, respectively.

From the Optimal CA to the Step CA.

We first quantify the gap between the two profits implied by the optimal CA (z^*) and the step CA (z^s) for each influence area j, defined as follows:

$$\mathsf{Gap}_{j,o2s} := \int_{x\in\mathcal{X}_j} \psi(z^*(x))dx - \int_{x\in\mathcal{X}_j} \psi(z^s(x))dx = \int_{x\in\mathcal{X}_j} \big(\psi(z^*(x)) - \psi(z^s(x))\big)dx.$$

Here we omit $W(x)$ and θ in function $\psi(\cdot)$ for brevity since the result holds for any $W(x)$ and θ.

Proof. Proof of Lemma 4. Recall we assume that $\psi(z)$ is twice differentiable and quasi-concave in z. Then, applying the Taylor expansion with the mean-value form of the remainder, there exists a $\bar{z}$ as a convex combination of z^s and z^* such that

$$\psi(z^s) = \psi(z^*) + \psi'(z^*)(z^s - z^*) + \frac{\psi''(\bar{z})}{2}(z^s - z^*)^2,$$

in which $\psi'(z^*) = 0$ due to the optimality of z^*. Subsequently,

$$\mathsf{Gap}_{j,o2s} = -\int_{x\in\mathcal{X}_j} \frac{\psi''(\bar{z}(x))}{2}(z^s(x) - z^*(x))^2 dx = -\int_{x\in\mathcal{X}_j} \frac{\psi''(\bar{z}(x))}{2\rho(x)}(z^s(x) - z^*(x))^2\rho(x)dx. \quad (17)$$

In the above integral, recall that $\rho(x)$ is the function of the spatial density distribution of customer demands, and that $z^s(x)$ is defined as a constant number for $x \in \mathcal{X}_j$ such that $z^s(x) = |\mathcal{X}_j|$. Therefore, $z^s(x)$ can be regarded as the mean of $z^*(x)$ over $\mathcal{X}_j$, i.e., $z^s(x) = \mathbb{E}[z^*(\mathcal{X}_j)] = \int_{x'\in\mathcal{X}_j} z^*(x')\rho(x')dx'$ for all $x \in \mathcal{X}_j$. We introduce the notation of variance of $z^*(x)$ over area $\mathcal{X}_j$ as $Var(z^*; \mathcal{X}_j) := \int_{x\in\mathcal{X}_j}(z^*(x) - \mathbb{E}[z^*(\mathcal{X}_j)])^2\rho(x)dx$. It follows that

$$Var(z^*; \mathcal{X}_j) = \int_{x\in\mathcal{X}_j} (z^*(x) - z^s(x))^2\rho(x)dx. \quad (18)$$

Moreover, $\frac{\psi''(\bar{z}(x))}{\rho(x)}$ represents the curvature (with respect to $\bar{z}(x)$) of the per-customer profit density function. In (19), we extend this point-wise curvature definition to be influence area-wise and assume that an upper bound of such a curvature exists as follows:

$$\frac{|\psi''(\bar{z}(x))|}{\rho(x')} \le \eta^\psi, \quad \forall x, x' \in \mathcal{X}_j,\ \bar{z}(x) \in \mathrm{Conv}\left(\{z^s(x), z^*(x)\}\right) \quad (19)$$

where $\eta^\psi > 0$ is a constant value, and $\mathrm{Conv}(\cdot)$ is a convex hull. Then substituting (18) and (19) into (17) completes the proof. □

From the Step CA to the Discrete Design.

We next quantify the gap between the two profits implied by the step CA and the discrete design for each influence area j, i.e.,

$$\mathsf{Gap}_{j,s2d} = \int_{x\in\mathcal{X}_j} \psi(z^s(x))dx - \int_{x\in\mathcal{X}_j} \varphi(A(z^*))dx = \int_{x\in\mathcal{X}_j} \big(\psi(z^s(x)) - \varphi(A(z^*))\big)dx.$$

In the operations of mobile retail stores, this gap stems from the disparities between continuous profit and discrete design profit. The CA recipe $z^s(x) = |\mathcal{X}_j|$ is derived from discretized decisions about store locations, compelling a detailed examination of both the discretization procedure and the

profit function. Assumption 1 offers analytical convenience that captures the essence implied by the gap between the step CA and the discrete design. Ouyang and Daganzo (2006) use Assumptions 1(I)–(III) and another strict assumption that $\rho(x)$ is a constant within $\mathcal{X}_j$ to prove that the cost implied by the CA recipe is a lower bound for that of the discrete implementation for problems without facility costs. In our analysis of the CA gap, our objective is to determine both its direction and magnitude, building upon the findings from Ouyang and Daganzo (2006). To achieve, we further impose the value caps in Assumption 2. Our quantification of this gap can be considered as a sharpening and extension of the result in Ouyang and Daganzo (2006), as stated in Lemma 5.

Proof of Lemma 5. The proof is based on the proof of the theorem in Ouyang and Daganzo (2006), with additional sharpening results. Before presenting the proof, we provide the following sharpened Jensen's inequality due to Liao and Berg (2019):

Lemma 10. Suppose that $\varphi(d)$ is a twice differentiable function of $d \in \mathcal{D}$ and that D is a one-dimensional random variable with variance $Var(D)$. Then

$$Var(D) \inf_{d\in\mathcal{D}} \frac{\varphi''(d)}{2} \leq \mathbb{E}[\varphi(D)] - \varphi(\mathbb{E}[D]) \leq Var(D) \sup_{d\in\mathcal{D}} \frac{\varphi''(d)}{2}.$$

Back to Lemma 5, recall that the profit density function φ consists of the revenue term, the inventory replenishment cost, and the facility cost, as expressed in (3) and (6) for the discrete and the CA models, respectively. We compare these two models component by component.

First notice that the total facility cost over the entire influence area $\mathcal{X}_j$, $\int_{\mathcal{X}_j} \varphi^f(x)dx$, depends only on the total daily sales in that area. The mean of daily sales in the step CA model is $\int_{\mathcal{X}_j} \rho(x)z^s(x)dx/|\mathcal{X}_j| = \int_{\mathcal{X}_j} \rho(x)dx$ (since $z^s(x) = |\mathcal{X}_j|$), which is equal to the daily sales in the discrete model. Therefore, by Assumption 1(II) that φ^f is an affine function, there is no gap in the facility cost between the step CA model and the discrete design. In other words,

$$\int_{\mathcal{X}_j} \varphi^f \left(\int_{\mathcal{X}_j} \rho(x)dx \right) dx - \int_{\mathcal{X}_j} \varphi^f \left(\rho(x)z^s(x)\right) dx = 0. \tag{20}$$

Following the same argument, the gap due to the revenue term is zero, too.

We next examine the inventory replenishment cost, which depends on both truck routing distance and the total daily sales. The gap in the inventory replenishment cost between these two models is decomposed into

$$\begin{aligned}
&\int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, \mathcal{X}_j \right) dx - \int_{\mathcal{X}_j} \varphi^i(\rho(x)z^s(x), z^s(x))dx \\
&= \left[\int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, \mathcal{X}_j \right) dx - \int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, z^s(x) \right) dx \right] \\
&\quad + \left[\int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, z^s(x) \right) dx - \int_{\mathcal{X}_j} \varphi^i(\rho(x)z^s(x), z^s(x))dx \right],
\end{aligned} \tag{21}$$

where the first term represents the difference in truck routing distances, and the second term represents the difference in daily sales in the area $\mathcal{X}_j$. As previously discussed, the CA approach in this cost segment involves replacing the restocking truck routing trip distance $d(x', x_j)$ with $d(x', x)$ for any $x' \in \mathcal{X}$ and $x \in \mathcal{X}_j$, thereby yielding the optimal routing distance $\beta_{\mathsf{TSP}} \int_{x\in\mathcal{X}_j} 1/\sqrt{z(x)}dx$. Since $d(x', x_j)$ is the average of $d(x', x)$ by Assumption 1(I), we obtain

$$0 \leq \int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, \mathcal{X}_j \right) dx - \int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x)dx, z^s(x) \right) dx \leq \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j). \tag{22}$$

Here the first inequality is due to Jensen's inequality and the concavity of φ^i in distance (Assumption 1(III)). The second inequality is due to the sharpened Jensen's inequality (i.e., Lemma 10) and Assumptions 2.

Note that $z^s(x) = |\mathcal{X}_j|$ implies that $\int_{\mathcal{X}_j} \rho(x) z^s(x) dx / |\mathcal{X}_j| = \int_{\mathcal{X}_j} \rho(x) dx$, which means the second term of (21) is the gap of taking average over daily sales. Similar to (22), we have the following inequalities:

$$0 \leq \int_{\mathcal{X}_j} \varphi^i \left(\int_{\mathcal{X}_j} \rho(x) dx, z^s(x) \right) dx - \int_{\mathcal{X}_j} \varphi^i \left(\rho(x) z^s(x), z^s(x) \right) dx \leq \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j). \tag{23}$$

Combining (20)–(23) completes the proof of Lemma 5(I). Lemma 5(II) can be similarly proved by applying the sharpened Jensen's inequality instead of Assumption 1(III) for the first inequality in (22) and (23). □

Closing the CA Gap.

Proof of Theorem 1. Suppose Assumptions 1 and 2 hold. Combining Lemmas 4 and 5(I) immediately yields the following bound for each influences zone j:

$$0 \leq \mathsf{Gap}_{j,o2s} + \mathsf{Gap}_{j,s2d} \leq \left(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) + \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) \right).$$

It follows that the CA gap is bounded as follows:

$$0 \leq \mathsf{Gap}_{\mathsf{CA}} \leq \sum_{j=1}^{N} \left(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) + \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) \right) \leq \beta_{\mathsf{CA}}.$$

Alternatively, if relaxing Assumption 1(III), combining Lemmas 4 and 5(II) yields:

$$\left(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) - \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) - \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) \right) \leq$$
$$\mathsf{Gap}_{j,o2s} + \mathsf{Gap}_{j,s2d} \leq \left(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) + \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) \right).$$

Summing over all influence areas, we can reach our conclusion that

$$|\mathsf{Gap}_{\mathsf{CA}}| \leq \beta_{\mathsf{CA}}.$$ □

C.2 Proofs in Section 5.2

Proof of Lemma 6. Recall that g_{sj} denotes the gradient $\nabla_\theta f_\theta(A_s; W_s, \mathcal{X}_{sj})$ and λ is the regularization parameter. Since $V_t = \sum_{s=1}^{t-1} \sum_{j=1}^{N_s} g_{sj} {g_{sj}}^\top + \lambda I \succeq \lambda I \succeq 0$, we have $V_t^{-1} \preceq \lambda^{-1} I$. Cauchy–Schwarz inequality implies

$$\begin{aligned}
\|\xi_t - \lambda\theta^*\|_{V_t^{-1}} &\leq \|\xi_t\|_{V_t^{-1}} + \|\lambda\theta^*\|_{V_t^{-1}} \\
&\leq \|\xi_t\|_{V_t^{-1}} + \sqrt{(\lambda\theta^*)^\top \lambda^{-1} I (\lambda\theta^*)} \\
&= \|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\|\theta^*\|_2 \\
&\leq \|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta.
\end{aligned}$$

Observe that (Oracle) gives

$$\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*) - \epsilon_{sj})g_{sj} + \lambda\hat{\theta}_t = 0,$$

which yields

$$\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}\epsilon_{sj}g_{sj} - \lambda\theta^* = \sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*))g_{sj} + \lambda(\hat{\theta}_t - \theta^*). \tag{24}$$

Thus we have

$$\begin{aligned}
&(\|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta)^2 \geq \|\xi_t - \lambda\theta^*\|^2_{V_t^{-1}} \\
&= \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(Y_{sj} - f_{sj}(\theta^*))g_{sj} - \lambda\theta^*\right)^\top \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)^{-1} \\
&\quad \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(Y_{sj} - f_{sj}(\theta^*))g_{sj} - \lambda\theta^*\right) \\
&\overset{(24)}{=} \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*))g_{sj} + \lambda\hat{\theta}_t - \lambda\theta^*\right)^\top \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)^{-1} \\
&\quad \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*))g_{sj} + \lambda\hat{\theta}_t - \lambda\theta^*\right).
\end{aligned}$$

Recall that the mean demand $f_{sj}(\theta) = \int_{\mathcal{X}_{sj}} \theta^\top \kappa(A_t, W_t(x))dx$ is a linear function in θ. Therefore, the gradient $\nabla_\theta f_{sj}(\theta)$ is a constant that does not depend on θ, and we have $\nabla_\theta f_{sj}(\theta) = g_{sj}$ holds for every θ. Applying the mean-value theorem, there exists $\bar{\theta}_t$ which is a convex combination of θ^* and $\hat{\theta}_t$ such that

$$f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*) = \nabla_{\bar{\theta}_t} f_{sj}(\bar{\theta}_t)^\top(\hat{\theta}_t - \theta^*) = g_{sj}^\top(\hat{\theta}_t - \theta^*),$$

which implies that

$$\begin{aligned}
&\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}(f_{sj}(\hat{\theta}_t) - f_{sj}(\theta^*))g_{sj} + \lambda(\hat{\theta}_t - \theta^*) \\
&= \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)(\hat{\theta}_t - \theta^*).
\end{aligned}$$

According to the symmetric property,

$$\begin{aligned}
&(\|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta)^2 \\
&\geq (\hat{\theta}_t - \theta^*)^\top \left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)\left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)^{-1}\left(\sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top + \lambda I\right)(\hat{\theta}_t - \theta^*) \\
&= \|\hat{\theta}_t - \theta^*\|^2_{V_t}.
\end{aligned}$$

□

Proof of Lemma 7. It follows similarly from Theorem 20.4 in Lattimore and Szepesvári (2020). The difference is that, in our setting, multiple observations are received at the same moment. For completeness, we provide the full proof here.

Let $U_t = \sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}^\top$ and $H = \lambda I \in \mathbb{R}^{d\times d}$ so $V_t = U_t + H$. We first prove that for all $x \in \mathbb{R}^d$ the process $M_t(x) = \exp(\langle x, \xi_t\rangle - \frac{\sigma^2}{2}\|x\|^2_{U_t})$ is an $\mathcal{H}$-adapted non-negative supermatingale with $M_0(x) \le 1$. We need to show that $\mathbb{E}[M_{t+1}(x)|\mathcal{F}_t] \le M_t(x)$ almost surely. Since ϵ_{tj} is conditionally σ-subgaussian, then for any $1 \le j \le N_t$, we have

$$\mathbb{E}\left[\exp\left(x^\top \epsilon_{tj} g_{tj} - \frac{\sigma^2}{2}\|x\|^2_{g_{tj}g_{tj}^\top}\right)\middle|\mathcal{H}_t\right] \le 1.$$

Hence

$$\begin{aligned}\mathbb{E}[M_{t+1}(x)|\mathcal{H}_t] &= \mathbb{E}\left[\exp\left(\langle x, \xi_{t+1}\rangle - \frac{\sigma^2}{2}\|x\|^2_{U_{t+1}}\right)\middle|\mathcal{H}_t\right] \\ &= M_t(x)\mathbb{E}\left[\exp\left(\sum_{j=1}^{N_t} x^\top \epsilon_{tj} g_{tj} - \frac{\sigma^2}{2}\|x\|^2_{\sum_{j=1}^{N_t} g_{tj}g_{tj}^\top}\right)\middle|\mathcal{H}_t\right] \\ &\le M_t(x) \text{ a.s.}\end{aligned}$$

Since $M_0(x) \le 1$, we reach the conclusion that $M_t(x)$ is a non-negative supermartingale.

Define $h = \mathcal{N}(0, (\sigma^2 H)^{-1})$. According to the "*sections*" lemma in (Kallenberg 1997, Lemma 1.28), $\int_{\mathbb{R}^d} M_t(x)dh(x)$ is $\mathcal{H}_t$-measurable and is also a non-negative supermartingale. Let

$$\begin{aligned}\bar{M}_t &= \int_{\mathbb{R}^d} M_t(x)dh(x) \\ &= \frac{\sigma^d}{\sqrt{(2\pi)^d \det(H^{-1})}} \int_{\mathbb{R}^d} \exp\left(\langle x, \xi_t\rangle - \frac{\sigma^2}{2}\|x\|^2_{U_t} - \frac{\sigma^2}{2}\|x\|^2_H\right) dx.\end{aligned}$$

Note that $\bar{M}_0 \le 1$ is immediate. By the maximal inequality for the supermartingale $\bar{M}_t$, we have

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} \log(\bar{M}_t) \ge \log\left(\frac{1}{\delta}\right)\right) = \mathbb{P}\left(\sup_{t\in\mathbb{N}} \bar{M}_t \ge \frac{1}{\delta}\right) \le \delta. \tag{25}$$

Now we turn to studying $\bar{M}_t$. In the definition of $\bar{M}_t$, we reformulate the polynomial within the integrand as follows

$$\begin{aligned}\langle x, \xi_t\rangle - \frac{\sigma^2}{2}\|x\|^2_{U_t} - \frac{\sigma^2}{2}\|x\|^2_H &= \frac{1}{2}\|\sigma^{-1}\xi_t\|^2_{(H+U_t)^{-1}} - \frac{1}{2}\|\sigma x - (H+U_t)^{-1}\sigma^{-1}\xi_t\|^2_{H+U_t} \\ &= \frac{1}{2}\|\sigma^{-1}\xi_t\|^2_{V_t^{-1}} - \frac{1}{2}\|\sigma x - V_t^{-1}\sigma^{-1}\xi_t\|^2_{V_t}.\end{aligned}$$

The first term $\frac{1}{2}\|\sigma^{-1}\xi_t\|^2_{V_t^{-1}}$ does not depend on x and can be moved outside of the integral. In such a way, the integration equals

$$\bar{M}_t = \frac{\sigma^d}{\sqrt{(2\pi)^d \det(H^{-1})}} \cdot \exp\left(\frac{1}{2}\|\sigma^{-1}\xi_t\|^2_{V_t^{-1}}\right) \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\sigma x - V_t^{-1}\sigma^{-1}\xi_t\|^2_{V_t}\right) dx,$$

in which only a quadratic Gaussian term is integrated. By multidimensional Gaussian integral, we have

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\sigma x - V_t^{-1}\sigma^{-1}\xi_t\|^2_{V_t}\right) dx = \sigma^{-d}\sqrt{\frac{(2\pi)^d}{\det(V_t)}},$$

which implies that

$$\bar{M}_t = \left(\frac{\det(H)}{\det(V_t)}\right)^{1/2} \exp\left(\frac{1}{2\sigma^2}\|\xi_t\|^2_{V_t^{-1}}\right).$$

Then substituting this expression in Equation (25), we reach the conclusion. □

Proof of Lemma 8. Lemma 6 provides that $\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \|\xi_t\|_{V_t^{-1}} + \sqrt{\lambda}\beta_\Theta$. From Lemma 7 that it holds with probability at least $1-\delta$ that, for all $t \in [T]$,

$$\|\xi_t\|_{V_t^{-1}} \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}.$$

Let $\tilde{\lambda}_1, \cdots, \tilde{\lambda}_d$ be the eigenvalues of V_t. Note that Assumption 4(V) implies $\|g_{sj}\|_2 \leq |\mathcal{X}_{sj}|\beta_\kappa$. By the inequality of arithmetic and geometric means,

$$\begin{aligned}
\frac{\det(V_t)}{\lambda^d} &= \frac{1}{\lambda^d}\prod_{j=1}^{d}\tilde{\lambda}_j \leq \frac{1}{\lambda^d}\left(\frac{1}{d}\sum_{j=1}^{d}\tilde{\lambda}_j\right)^d = \left(\frac{\mathrm{tr}\,(V_t)}{\lambda d}\right)^d \\
&= \left(\frac{\mathrm{tr}\left(\lambda I + \sum_{s=1}^{t-1}\sum_{j=1}^{N_s} g_{sj}g_{sj}{}^\top\right)}{\lambda d}\right)^d \\
&= \left(1 + \frac{\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}\|g_{sj}\|_2^2}{\lambda d}\right)^d \\
&\leq \left(1 + \frac{\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N_s}\|g_{sj}\|_2\right)^2}{\lambda d}\right)^d \\
&\leq \left(1 + \frac{\sum_{s=1}^{t-1}\left(\sum_{j=1}^{N_s}|\mathcal{X}_{sj}|\beta_\kappa\right)^2}{\lambda d}\right)^d \\
&= \left(1 + \frac{(t-1)|\mathcal{X}|^2\beta_\kappa^2}{\lambda d}\right)^d.
\end{aligned}$$

Therefore, we conclude that with probability at least $1-\delta$, we have

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t,$$

where $\gamma_t = \sqrt{\lambda}\beta_\Theta + \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{(t-1)|\mathcal{X}|^2\beta_\kappa^2}{\lambda d}\right)}$. □

Proof of Lemma 9. Suppose the action A partitions the region $\mathcal{X}$ into a set of N influence areas, i.e., $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_N\}$. Fix j and we prove for $\mathcal{X}_j$. Let ϱ_j be the brevity of the daily sales handled by a store serving $\mathcal{X}_j$, which is used an argument of φ^i and φ^f in the following assumption; accordingly, $\varrho_j = \int_{\mathcal{X}_j}\rho_\theta(A, x; W(x))dx = f_\theta(A; W, \mathcal{X}_j)$ in discrete model and $\varrho = \rho(x)z(x)$ in CA model.

For notation simplicity, we use $\rho_\theta(x)$ as the brevity of $\rho_\theta(A, x; W(x))$ and φ_θ^i as the brevity of $\varphi^i\left(\int_{\mathcal{X}_j}\rho_\theta(x)dx, \mathcal{X}_j; W(x)\right)$ in the proof of this lemma, with φ_θ^f being similar abbreviation.

$$\begin{aligned}
|r_\theta(A; W, \mathcal{X}_j) - r_{\theta^*}(A; W, \mathcal{X}_j)| \leq &\left|\int_{\mathcal{X}_j}\bar{r}\rho_\theta(x)dx - \int_{\mathcal{X}_j}\bar{r}\rho_{\theta^*}(x)dx\right| + \left|\left(\int_{\mathcal{X}_j}\varphi_\theta^i dx - \int_{\mathcal{X}_j}\varphi_{\theta^*}^i dx\right)\right| \\
&+ \left|\left(\int_{\mathcal{X}_j}\varphi_\theta^f dx - \int_{\mathcal{X}_j}\varphi_{\theta^*}^f dx\right)\right|
\end{aligned}$$

where the gap on gross revenue can be simply rewritten as $\bar{r}|f_\theta(A; W, \mathcal{X}) - f_{\theta^*}(A; W, \mathcal{X})|$.

By Lipschitz continuity, the gap on inbound cost can be bounded as

$$\begin{aligned}\left|\left(\int_{\mathcal{X}_j}\varphi^i_\theta dx-\int_{\mathcal{X}_j}\varphi^i_{\theta^*}dx\right)\right| &\le \int_{\mathcal{X}_j}\alpha_i\left|\left(\int_{\mathcal{X}_j}\rho_\theta(x)dx-\int_{\mathcal{X}_j}\rho_{\theta^*}(x)dx\right)\right|dy\\ &= \alpha_i|\mathcal{X}_j||f_\theta(A;W,\mathcal{X}_j)-f_{\theta^*}(A;W,\mathcal{X}_j)|.\end{aligned}$$

The following inequality can be similarly obtained

$$\left|\left(\int_{\mathcal{X}_j}\varphi^f_\theta dx-\int_{\mathcal{X}_j}\varphi^f_{\theta^*}dx\right)\right| \le \alpha_f|\mathcal{X}_j||f_\theta(A;W,\mathcal{X}_j)-f_{\theta^*}(A;W,\mathcal{X}_j)|.$$

Thus we conclude that $L_r=\bar{r}+|\mathcal{X}|\alpha_i+|\mathcal{X}|\alpha_f$. □

Finally, our proof of the regret bound will also depend on a technical result, which we call the Elliptical Potential Lemma.

Lemma 11 (Elliptical potential lemma)**.**

$$\sum_{t=t_0+1}^{T}\min\left\{\sum_{j=1}^{N_t}\|g_{tj}\|^2_{V_t^{-1}},1\right\}\le 2d\log\left(\frac{d\lambda+T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right).$$

Proof of Lemma 11. Recall that $V_t=\sum_{s=1}^{t-1}\sum_{j=1}^{N_s}g_{sj}g_{sj}^\top+\lambda I$. Then we have

$$V_{t+1}=V_t+\sum_{j=1}^{N_t}g_{tj}g_{tj}^\top=V_t^{1/2}\Big(I+\sum_{j=1}^{N_t}V_t^{-1/2}g_{tj}g_{tj}^\top V_t^{-1/2}\Big)V_t^{1/2}.$$

According to the definition of V_t, we have

$$\begin{aligned}\det(V_{t+1}) &= \det\big(V_t+\sum_{j=1}^{N_t}g_{tj}g_{tj}^\top\big)\\ &= \det(V_t)\det\left(I+\sum_{j=1}^{N_t}V_t^{-1/2}g_{tj}g_{tj}^\top V_t^{-1/2}\right).\end{aligned}$$

Let $\lambda_1,\cdots,\lambda_d$ be the eigenvalues of $\sum_{j=1}^{N_t}u_{tj}u_{tj}^\top$ where $u_{tj}=V_t^{-1/2}g_{tj}$. Note that the eigenvalues of matrix $I+\sum_{j=1}^{N_t}u_{tj}u_{tj}^\top$ are $(1+\lambda_j)$ for $j=1,\cdots,d$. Then we have

$$\det\left(I+\sum_{j=1}^{N_t}u_{tj}u_{tj}^\top\right)=\prod_{j=1}^{d}(1+\lambda_j)\ge 1+\sum_{j=1}^{d}\lambda_j=1+\mathrm{tr}\left(\sum_{j=1}^{N_t}u_{tj}u_{tj}^\top\right)=1+\sum_{j=1}^{N_t}\|u_{tj}\|_2^2.$$

Therefore, we have the inequality

$$\det(V_{t+1})\ge\det(V_t)\left(1+\sum_{j=1}^{N_t}\|g_{tj}\|^2_{V_t^{-1}}\right).$$

Using that for any $x \geq 0$, $\min\{x, 1\} \leq 2\log(1+x)$, we get

$$\begin{aligned}
&\sum_{t=t_0+1}^{T} \min\left\{\sum_{j=1}^{N_t} \|g_{tj}\|^2_{V_t^{-1}}, 1\right\} \\
\leq& 2\sum_{t=t_0+1}^{T} \log\left(1+\sum_{j=1}^{N_t} \|g_{tj}\|^2_{V_t^{-1}}\right) \leq 2\log\left(\frac{\det(V_{T+1})}{\det(V_{t_0+1})}\right) \leq 2\log\left(\frac{\det(V_{T+1})}{\lambda^d_{\min}(V_{t_0+1})}\right) \\
\leq& 2\log\left(\frac{\det(V_{T+1})}{\lambda^d}\right) \leq 2d\log\left(\frac{d\lambda + T|\mathcal{X}|^2\beta^2_\kappa}{d\lambda}\right).
\end{aligned}$$

□

Proof of Theorem 2. According to Theorem 1, we have

$$|r_\theta(A(z_t^*); W_t) - r^\psi_\theta(z_t^*; W_t)| \leq \beta_{\mathsf{CA}} \text{ for all } 1 \leq t \leq T.$$

Define event $\mathcal{E}_t = \{\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t\}$. Under event $\mathcal{E}_t$, θ^* is contained in the uncertainty set $\Theta_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t} \leq \gamma_t\}$. Since θ_t is the optimal optimistic solution over the set Θ_t, we have

$$r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) \leq r^\psi_{\theta_t}(z_t^*(\cdot;\theta_t); W_t). \tag{26}$$

Thus, we can bound one-step regret as

$$\begin{aligned}
&r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t) \\
=&r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) + r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t) \\
\overset{(26)}{\leq}& r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) + r^\psi_{\theta_t}(z_t^*(\cdot;\theta_t); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t) \\
=&r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) + r^\psi_{\theta_t}(z_t^*(\cdot;\theta_t); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t) \\
&+ r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) - r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) \\
=&\Big(r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t)\Big) \\
&- \Big(r^\psi_{\theta^*}(z_t^*(\cdot;\theta^*); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t\Big) + \Big(r^\psi_{\theta_t}(z_t^*(\cdot;\theta_t); W_t) - r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t\Big).
\end{aligned}$$

Theorem 1 regarding the CA gap implies that

$$\begin{aligned}
&(r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t))\,\mathbf{1}(\mathcal{E}_t) \\
\leq& |r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t)|\,\mathbf{1}(\mathcal{E}_t) + 2\beta_{\mathsf{CA}}.
\end{aligned} \tag{27}$$

According to Lemma 9, we have

$$\begin{aligned}
&|r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t)| \\
\leq& \sum_{j=1}^{N_t} |r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj}) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj})| \\
\leq& L_r \sum_{j=1}^{N_t} |f_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj}) - f_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj})|.
\end{aligned} \tag{28}$$

Applying mean-value Theorem, there exists $\bar{\theta}_{tj}$ which is a convex combination of θ^* and θ_t such that

$$\begin{aligned}
&|f_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj}) - f_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj})| \\
=&\left|(\theta_t - \theta^*)^\top \nabla f_{\bar{\theta}_{tj}}(A(z_t^*(\cdot;\theta_t)); W_t, \mathcal{X}_{tj})\right|
\end{aligned}$$

$$
\begin{aligned}
&= \left| (\theta_t - \theta^*)^\top \int_{\mathcal{X}_{tj}} \kappa(A_t, W_t(x)) dx \right| \\
&= \left| (\theta_t - \theta^*)^\top g_{tj} \right| \\
&\leq \|\theta_t - \theta^*\|_{V_t} \|g_{tj}\|_{V_t^{-1}}, \qquad (29)
\end{aligned}
$$

where the last inequality is obtained from the Cauchy-Schwarz inequality.

Note that under event $\mathcal{E}_t = \{\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t\}$, we have $\|\theta_t - \theta^*\|_{V_t} \leq \|\theta_t - \hat{\theta}_t\|_{V_t} + \|\hat{\theta}_t - \theta^*\|_{V_t} \leq 2\gamma_t$. Combining inequalities (27)–(29) immediately yields

$$
\begin{aligned}
&(r_{\theta^*}(A(z_t^*(\cdot;\theta^*)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t))\mathbf{1}(\mathcal{E}_t) \\
\leq &|r_{\theta_t}(A(z_t^*(\cdot;\theta_t)); W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t)|\, \mathbf{1}(\mathcal{E}_t) + 2\beta_{\mathsf{CA}} \\
\leq &L_r \|\theta^* - \theta_t\|_{V_t} \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}} \mathbf{1}(\mathcal{E}_t) + 2\beta_{\mathsf{CA}} \\
\leq &2L_r \gamma_t \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}} + 2\beta_{\mathsf{CA}}.
\end{aligned}
$$

According to Assumption 4(I), the maximal profit is bounded by $r_{\max}$. Thus, we can bound the regret by

$$
\begin{aligned}
\mathsf{Regret}_\pi(T) \cdot \mathbf{1}\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right) \leq &r_{\max} t_0 + \sum_{t=t_0+1}^{T} \mathbb{E}_\pi[(r_{\theta^*}(A_t^*; W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t)); W_t))\mathbf{1}(\mathcal{E}_t)] \\
\leq &r_{\max} t_0 + 2L_r \gamma_T \sum_{t=t_0+1}^{T} r_{\max} \wedge \left(\sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}\right) + 2\beta_{\mathsf{CA}} T.
\end{aligned}
$$

Note that the equation $1 \wedge x = \sqrt{1 \wedge x^2}$ holds for every $x > 0$, and the Cauchy-Schwarz inequality yields $\left(\sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}\right)^2 \leq N_t \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}^2$. Since we assume $r_{\max} > 1$, it follows that

$$
\begin{aligned}
&\mathsf{Regret}_\pi(T) \cdot \mathbf{1}\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right) \\
\leq &r_{\max} t_0 + 2r_{\max} L_r \gamma_T \sum_{t=t_0+1}^{T} \sqrt{1 \wedge \left(\sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}\right)^2} + 2\beta_{\mathsf{CA}} T \\
\leq &r_{\max} t_0 + 2r_{\max} L_r \gamma_T \sum_{t=t_0+1}^{T} \sqrt{N_t} \sqrt{1 \wedge \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}^2} + 2\beta_{\mathsf{CA}} T \\
\leq &r_{\max} t_0 + 2r_{\max} L_r \gamma_T \sqrt{\left(\sum_{t=t_0+1}^{T} N_t\right)\left(\sum_{t=t_0+1}^{T} 1 \wedge \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}}^2\right)} + 2\beta_{\mathsf{CA}} T \\
\leq &r_{\max} t_0 + 2r_{\max} L_r \gamma_T \sqrt{2N_{\max} dT \log\left(\frac{d\lambda + T|\mathcal{X}|^2 \beta_\kappa^2}{d\lambda}\right)} + 2\beta_{\mathsf{CA}} T,
\end{aligned}
$$

where the last inequality holds according to Lemma 11.

Lemma 8 states that the event $\bigcap_{t=t_0+1}^{T} \mathcal{E}_t$ occurs with probability at least $1-\delta$. The probability of the event $\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right)^c$ is no greater than δ. Thus, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\mathsf{Regret}_\pi(T) > r_{\max} t_0 + 2r_{\max} L_r \gamma_T \sqrt{2N_{\max} dT \log\left(\frac{d\lambda + T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\beta_{\mathsf{CA}} T\right) \\
\leq &\mathbb{E}\left[\mathbf{1}\left(\mathsf{Regret}_\pi(T)\mathbf{1}\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right) > r_{\max} t_0 \right.\right. \\
&\left.\left. + 2r_{\max} L_r \gamma_T \sqrt{2N_{\max} dT \log\left(\frac{d\lambda + T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\beta_{\mathsf{CA}} T\right)\right] + \mathbb{P}\left(\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right)^c\right) \\
= &\mathbb{E}[0] + \mathbb{P}\left(\left(\bigcap_{t=t_0+1}^{T} \mathcal{E}_t\right)^c\right) \\
\leq &\delta.
\end{aligned}$$

□

Lemma 12. Set $t_0^F = \max\left\{\frac{\sqrt{T}}{(1-\delta)\underline{\lambda}}, \frac{(\log(d)-\log(\delta))|\mathcal{X}|^2\beta_\kappa^2}{(\delta+(1-\delta)\log(1-\delta))\underline{\lambda}}\right\}$. Suppose Assumption 5 holds. If using policy π for t_0^F time periods, then it holds that

$$\mathbb{P}\left(\lambda_{\min}\left(V_{t_0^F+1}\right) \leq \sqrt{T}\right) \leq \delta.$$

Proof of Lemma 12. Define $Z_s = \sum_{j=1}^{N_s(A)} \int_{\mathcal{X}_{sj}} \kappa(A, x; W_s(x))dx \int_{\mathcal{X}_{sj}} \kappa(A, x; W_s(x))^\top dx$. For any randomized action A, an upper bound for the eigenvalues of Z_s is given by

$$\lambda_{\max}(Z_s) \leq \sum_{i=1}^{d} \lambda_i(Z_s) = \mathrm{tr}(Z_s) = \mathrm{tr}\left(\sum_{j=1}^{N_s} g_{sj} g_{sj}^\top\right) = \sum_{j=1}^{N_s} \mathrm{tr}(g_{sj}^\top g_{sj}) \leq |\mathcal{X}|^2\beta_\kappa^2 \text{ almost surely.}$$

Recall $V_t = \sum_{s=1}^{t-1} Z_s + \lambda I$. From Assumption 5, under policy π, we have

$$\lambda\left(\mathbb{E}\left[V_{t_0^F+1}\right]\right) = \lambda\left(\sum_{s=1}^{t_0^F} \mathbb{E}\left[Z_s\right] + \lambda I\right) \geq \underline{\lambda} t_0^F + \lambda.$$

Let $\mu_{\min} = \underline{\lambda} t_0^F + \lambda$, the minimum eigenvalue of $\mathbb{E}[V_{t_0^F+1}]$. Applying Lemma 13 (Tropp (2012, Theorem 1.1)), we can bound $\lambda_{\min}(V_{t_0^F+1})$ by

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}\left(V_{t_0^F+1}\right) \leq \sqrt{T}\right) &\leq \mathbb{P}\left(\lambda_{\min}\left(V_{t_0^F+1}\right) \leq (1-\delta)\underline{\lambda} t_0^F\right) \\
&\leq \mathbb{P}\left(\lambda_{\min}\left(V_{t_0^F+1}\right) \leq (1-\delta)\mu_{\min}\right) \\
&\leq d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\frac{\mu_{\min}}{|\mathcal{X}|^2\beta_\kappa^2}} \\
&\leq \delta,
\end{aligned}$$

where the first inequality is because $t_0^F \geq \frac{\sqrt{T}}{(1-\delta)\underline{\lambda}}$, and the last inequality is because

$$t_0^F \geq \frac{(\log(d) - \log(\delta))|\mathcal{X}|^2\beta_\kappa^2}{(\delta + (1-\delta)\log(1-\delta))\underline{\lambda}}.$$

Thus, we reach our conclusion. □

Proof of Theorem 3. Recall that we solve the maximization problem (12) in order to obtain θ_t. Therefore, θ_t satisfies that for any $\theta \in \Theta_t$, it holds that

$$r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta_t - \hat{\theta}_t) \geq r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta - \hat{\theta}_t). \quad (30)$$

Applying the mean-value theorem, there exists $\bar{\theta}_t$ which is a convex combination of $\hat{\theta}_t$ and θ^* such that

$$r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) = r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t), \quad (31)$$

and there exists $\bar{\theta}'_t$ which is a convex combination of $\hat{\theta}_t$ and θ_t such that

$$r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) = r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t) + \nabla r^{\psi}_{\bar{\theta}'_t}(z^*_t(\cdot;\bar{\theta}'_t);W_t)^\top(\hat{\theta}_t - \theta_t). \quad (32)$$

Then when event $\mathcal{E}_t$ holds, we have

$$\begin{aligned}
&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
=&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) + r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
\overset{(31)}{=}&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) + r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t)\\
&- r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
=&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) + r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t)\\
&+ \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
\overset{(30)}{\leq}&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t) + r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) + \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta_t - \hat{\theta}_t)\\
&+ \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
&+ r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t)) + r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t)) - r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t),
\end{aligned}$$

where the last inequality is because $\theta^* \in \Theta_t$ under event $\mathcal{E}_t$. By regrouping these terms in the above inequality, we further have

$$\begin{aligned}
&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
\leq&\Big(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\Big) + \Big(r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r^{\psi}_{\theta^*}(z^*_t(\cdot;\theta^*);W_t)\Big)\\
&+ \Big(r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t)) - r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t)\Big) + \Big(r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) - r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t)\Big)\\
&+ \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta_t - \hat{\theta}_t) + \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t).
\end{aligned}$$

Theorem 1 regarding the CA gap implies that

$$\begin{aligned}
&r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\\
\leq&\Big(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\Big) + 2\beta_{\mathsf{CA}}\\
&+ \Big(r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t) - r^{\psi}_{\theta_t}(z^*_t(\cdot;\theta_t);W_t)\Big) + \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta_t - \hat{\theta}_t)\\
&+ \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t)\\
\overset{(32)}{=}&\Big(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\Big) + 2\beta_{\mathsf{CA}}\\
&+ \nabla r^{\psi}_{\bar{\theta}'_t}(z^*_t(\cdot;\bar{\theta}'_t);W_t)^\top(\hat{\theta}_t - \theta_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\hat{\theta}_t - \theta_t)\\
&+ \nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t)^\top(\theta^* - \hat{\theta}_t)\\
=&\Big(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t)\Big) + 2\beta_{\mathsf{CA}}
\end{aligned}$$

$$+ (\nabla r^{\psi}_{\bar{\theta}'_t}(z^*_t(\cdot;\bar{\theta}'_t);W_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t))^\top(\hat{\theta}_t - \theta_t)$$
$$+ (\nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t)^\top - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t))^\top(\theta^* - \hat{\theta}_t).$$

Applying the mean-value theorem, there exists $\tilde{\theta}'_t$ which is a convex combination of $\bar{\theta}'_t$ and $\hat{\theta}_t$ such that

$$\begin{aligned}
&(\nabla r^{\psi}_{\bar{\theta}'_t}(z^*_t(\cdot;\bar{\theta}'_t);W_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t))^\top(\hat{\theta}_t - \theta_t)\\
=&(\bar{\theta}'_t - \hat{\theta}_t)^\top\nabla^2 r^{\psi}_{\tilde{\theta}'_t}(z^*_t(\cdot;\tilde{\theta}'_t);W_t)(\hat{\theta}_t - \theta_t)\\
\le&\hbar_f\|\hat{\theta}_t - \theta_t\|_2^2,
\end{aligned}$$

where the inequality is due to the Cauchy-Schwarz inequality and our Assumption 4(II) that $\hbar_f$ is an upper bound for the Euclidean norm of gradient $\nabla^2 r^{\psi}_{\tilde{\theta}'_t}(z^*_t(\cdot;\tilde{\theta}'_t);W_t)$. Similarly, there exists $\tilde{\theta}_t$ which is a convex combination of $\bar{\theta}_t$ and $\hat{\theta}_t$ such that

$$\begin{aligned}
&(\nabla r^{\psi}_{\bar{\theta}_t}(z^*_t(\cdot;\bar{\theta}_t);W_t) - \nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot;\hat{\theta}_t);W_t))^\top(\theta^* - \hat{\theta}_t)\\
=&(\bar{\theta}_t - \hat{\theta}_t)^\top\nabla^2 r^{\psi}_{\tilde{\theta}_t}(z^*_t(\cdot;\tilde{\theta}_t);W_t)(\theta^* - \hat{\theta}_t)\\
\le&\hbar_f\|\theta^* - \hat{\theta}_t\|_2^2.
\end{aligned}$$

Next, we need to bound the norms $\|\hat{\theta}_t - \theta_t\|_2^2$ and $\|\theta^* - \hat{\theta}_t\|_2^2$. Let $\lambda_{\min}(V_t)$ denote the minimum eigenvalue of the matrix V_t. Since $\|\hat{\theta}_t - \theta_t\|_{V_t} \le \gamma_t$ holds by our constructed uncertainty set, we have

$$\begin{aligned}
\|\hat{\theta}_t - \theta_t\|_2^2 &= \frac{1}{\lambda_{\min}(V_t)}(\hat{\theta}_t - \theta_t)^\top(\lambda_{\min}(V_t)I)(\hat{\theta}_t - \theta_t)\\
&\le \frac{1}{\lambda_{\min}(V_t)}(\hat{\theta}_t - \theta_t)^\top V_t(\hat{\theta}_t - \theta_t)\\
&= \frac{1}{\lambda_{\min}(V_t)}\|\hat{\theta}_t - \theta_t\|_{V_t}^2\\
&\le \frac{1}{\lambda_{\min}(V_t)}\gamma_t^2.
\end{aligned}$$

Similarly, under the event $\mathcal{E}_t = \{\|\hat{\theta}_t - \theta^*\|_{V_t} \le \gamma_t\}$, we also have $\|\theta^* - \hat{\theta}_t\|_2^2 \le \frac{1}{\lambda_{\min}(V_t)}\gamma_t^2$.

Therefore, we can bound the one-step regret by

$$\begin{aligned}
&(r_{\theta^*}(A(z^*_t(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t)\\
\le&(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t) + \hbar_f\|\hat{\theta}_t - \theta_t\|_2^2 + \hbar_f\|\theta^* - \hat{\theta}_t\|_2^2 + 2\beta_{\mathsf{CA}})\mathbf{1}(\mathcal{E}_t)\\
\le&\left(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t) + \frac{2\hbar_f\gamma_t^2}{\lambda_{\min}(V_t)} + 2\beta_{\mathsf{CA}}\right)\mathbf{1}(\mathcal{E}_t). \qquad (33)
\end{aligned}$$

In the proof of Theorem 2, we have already shown that

$$\begin{aligned}
&(r_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t) - r_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t)\\
\le&L_r\sum_{j=1}^{N_t}|f_{\theta_t}(A(z^*_t(\cdot;\theta_t));W_t,\mathcal{X}_{tj}) - f_{\theta^*}(A(z^*_t(\cdot;\theta_t));W_t,\mathcal{X}_{tj})|\,\mathbf{1}(\mathcal{E}_t)\\
\le&L_r\|\theta^* - \theta_t\|_{V_t}\sum_{j=1}^{N_t}\|g_{tj}\|_{V_t^{-1}}\mathbf{1}(\mathcal{E}_t)\\
\le&2L_r\gamma_t\sum_{j=1}^{N_t}\|g_{tj}\|_{V_t^{-1}}.
\end{aligned}$$

It follows that one-step regret in (33) is bounded by

$$\begin{aligned}&(r_{\theta^*}(A(z_t^*(\cdot;\theta^*));W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t)\\ \leq& 2L_r\gamma_t \sum_{j=1}^{N_t} \|g_{tj}\|_{V_t^{-1}} + \frac{2\hbar_f\gamma_t^2}{\lambda_{\min}(V_t)} + 2\beta_{\mathsf{CA}}.\end{aligned}$$

Recall that $\mathcal{E}_t = \{\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \gamma_t\}$ and we define $\mathcal{E}^\lambda = \{\lambda(V_{t_0^F}) \geq \sqrt{T}\}$. According to Lemma 12, it holds that $\mathbb{P}(\mathcal{E}^\lambda) \geq 1-\delta$. Therefore, we can bound the regret by

$$\begin{aligned}&\mathsf{Regret}_\pi(T)\mathbf{1}\left(\bigcap_{t=t_0^F+1}^{T} \mathcal{E}_t \cap \mathcal{E}^\lambda\right)\\ \leq& r_{\max}t_0^F + \sum_{t=t_0^F+1}^{T} \mathbb{E}_\pi[(r_{\theta^*}(A_t^*;W_t) - r_{\theta^*}(A(z_t^*(\cdot;\theta_t));W_t))\mathbf{1}(\mathcal{E}_t\cap\mathcal{E}^\lambda)]\\ \leq& r_{\max}t_0^F + 2L_r\gamma_T \sum_{t=t_0^F+1}^{T} r_{\max} \wedge \left(\sum_{j=1}^{N_t}\|g_{tj}\|_{V_t^{-1}}\right) + \sum_{t=t_0^F+1}^{T} 2\hbar_f\cdot\gamma_t^2\frac{1}{\lambda_{\min}(V_t)}\mathbf{1}(\mathcal{E}^\lambda) + 2\beta_{\mathsf{CA}}T\\ \leq& r_{\max}t_0^F + 2L_r\gamma_T \sum_{t=t_0^F+1}^{T} r_{\max} \wedge \left(\sum_{j=1}^{N_t}\|g_{tj}\|_{V_t^{-1}}\right) + 2\hbar_f\gamma_T^2 \sum_{t=t_0^F+1}^{T}\frac{1}{\sqrt{T}} + 2\beta_{\mathsf{CA}}T\\ \leq& r_{\max}t_0^F + 2r_{\max}L_r\gamma_T\sqrt{2N_{\max}dT\log\left(\frac{d\lambda+T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\hbar_f\gamma_T^2\sqrt{T} + 2\beta_{\mathsf{CA}}T,\end{aligned}$$

where the last inequality holds according to Lemma 11, similar to the proof of Theorem 2.

Since Lemma 8 implies $\mathbb{P}\left(\left(\bigcap_{t=t_0^F+1}^{T}\mathcal{E}_t\right)^c\right) \leq \delta$, and Lemma 12 implies $\mathbb{P}((\mathcal{E}^\lambda)^c) \leq \delta$, we conclude that

$$\begin{aligned}&\mathbb{P}\left(\mathsf{Regret}_\pi(T) \geq r_{\max}t_0^F + 2r_{\max}L_r\gamma_T\sqrt{2N_{\max}dT\log\left(\frac{d\lambda+T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)} + 2\hbar_f\gamma_T^2\sqrt{T} + 2\beta_{\mathsf{CA}}T\right)\\ \leq& \mathbb{E}\left[\mathbf{1}\left(\mathsf{Regret}_\pi(T)\mathbf{1}\left(\bigcap_{t=t_0^F+1}^{T}\mathcal{E}_t\cap\mathcal{E}^\lambda\right) \geq r_{\max}t_0^F + 2r_{\max}L_r\gamma_T\sqrt{2N_{\max}dT\log\left(\frac{d\lambda+T|\mathcal{X}|^2\beta_\kappa^2}{d\lambda}\right)}\right.\right.\\ &\left.\left. + 2\hbar_f\gamma_T^2\sqrt{T} + 2\beta_{\mathsf{CA}}T\right)\right] + \mathbb{P}\left(\left(\bigcap_{t=t_0^F+1}^{T}\mathcal{E}_t\right)^c\right) + \mathbb{P}((\mathcal{E}^\lambda)^c)\\ =& \mathbb{E}[0] + \mathbb{P}\left(\left(\bigcap_{t=t_0^F+1}^{T}\mathcal{E}_t\right)^c\right) + \mathbb{P}((\mathcal{E}^\lambda)^c)\\ \leq& 2\delta.\end{aligned}$$

□

D Supplementary results

D.1 Supplementary lemmas

Lemma 13 (Theorem 1.1 in Tropp (2012))**.** Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies

$$\mathbf{X}_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \ \text{ almost surely.}$$

Define

$$\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}[\mathbf{X}_k]\right) \quad \text{and} \quad \mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}[\mathbf{X}_k]\right).$$

Then

$$\mathbb{P}\left(\lambda_{min}\left(\sum_k \mathbf{X}_k\right) \le (1-\delta)\mu_{\min}\right) \le d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1), \text{ and}$$

$$\mathbb{P}\left(\lambda_{max}\left(\sum_k \mathbf{X}_k\right) \ge (1+\delta)\mu_{\max}\right) \le d\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \text{for } \delta \ge 0.$$

E Extensions

There are several extensions of the mobile retail problem that are worth further exploring. In this section, we discuss the effect of one-to-one inventory replenishment, delivering products to customers. In addition, we analyze the online location decision for last-mile delivery with micro-depots.

E.1 One-to-one inventory replenishment

In the three cases in our paper, we assume there is a truck visiting multiple stores to replenish the inventory. To extend the analysis, we replace the assumption of one-to-many replenishment in Case 3 by an one-to-one setting. Namely, a truck loads products from a distribution center at x_d and visits one store at a time for restocking. There is fixed cost for replenishment and we denote it by a_i. As discussed in Section 3.3, for a store located at x_{tj} and serving the area $\mathcal{X}_{tj}$, the replenishment frequency is $\int_{x\in\mathcal{X}_{tj}} \rho_\theta(x)dx/S$. A truck incurs cost c_t per kilometer of travel and travels $2d(x_d, x_{tj})$ kilometer for each replenishment. Therefore, the daily inventory replenishment cost for that store is given by $a^i + 2c_t d(x_d, x_{tj}) \int_{x\in\mathcal{X}_{tj}} \rho_\theta(x)dx/S$. The CA of cost density function for replenishment is given as follows:

$$\varphi^i\Big(d(x_d,x), \rho_\theta(x)z_t(x), z_t(x); W_t(x)\Big) = \frac{a^i + 2c_t d(x_d,x)\frac{\rho_\theta(x)z_t(x)}{S}}{z_t(x)} = \frac{a^i}{z_t(x)} + 2\frac{c_t}{S}d(x_d,x)\rho_\theta(x).$$

Similar to Case 3, the resulting CA of profit function is

$$\begin{aligned} r^{\psi}_{\theta}(z_t; W_t) &= \int_{x\in\mathcal{X}_t} \psi_\theta(z_t(x); W_t(x))dx \\ &= \int_{x\in\mathcal{X}_t} \left(\left(\bar{r} - a^f - 2\frac{c_t}{S}d(x_d,x)\right)\theta^\top W_t(x) \exp\left\{-c_0\frac{2}{3\sqrt{\pi}}\sqrt{z_t(x)}\right\} - \frac{a^i+b^f}{z_t(x)}\right)dx. \end{aligned}$$

At each $x \in \mathcal{X}_t$, we can evaluate the optimal solution $z_t^*(x;\theta)$ by first-order condition

$$\begin{aligned} z_t^*(x;\theta) &\in \left\{ z \,\middle|\, \frac{\partial \psi_\theta}{\partial z}(z; W_t(x)) = 0 \right\} \\ &= \left\{ z \,\middle|\, \left(\bar{r} - a^f - 2\frac{c_t}{S}d(x_d,x)\right)\theta^\top W_t(x) \exp\left\{-c_0\frac{2}{3\sqrt{\pi}}\sqrt{z}\right\} - \frac{3\sqrt{\pi}(a^i+b^f)}{c_0 z^{\frac{3}{2}}} = 0 \right\}. \end{aligned}$$

Specifically, the optimal $z_t^*(x;\theta)$ has a closed-form solution

$$z_t^*(x;\theta) = \frac{81\pi}{4c_0^2}\left(W_0\left(-\frac{2}{3}\left(\frac{(a^i+b^f)c_0^2}{9\pi(\bar{r} - a^f - 2\frac{c_t}{S}d(x_d,x))\theta^\top W_t(x)}\right)^{\frac{1}{3}}\right)\right)^2,$$

where $W_0(\cdot)$ is the principal branch of Lambert W function.

Despite the closed-form $z^*(x;\theta)$, the maximization problem (OFL-CA) is still intricate to solve. Therefore, we can choose to apply the Algorithm CA-O Faster Learning to easily solve the online store location problem. Lemma 2 provides the closed-form expression of gradient of CA profit function as

$$\nabla r^{\psi}_{\hat{\theta}_t}(z^*_t(\cdot,\hat{\theta}_t);W_t) = \int_{x\in\mathcal{X}_t}\left(\bar{r}-a^f-2\frac{c_t}{S}d(x_d,x)\right)\exp\left\{-c_0\frac{2}{3\sqrt{\pi}}\sqrt{z^*_t(x;\hat{\theta}_t)}\right\}W_t(x)dx.$$

Given that Assumptions 1-4 remain valid in this one-to-one inventory replenishment setting, Theorem 3 still provides an evaluation of the regret performance of Algorithm CA-O Faster Learning.

E.2 Delivery to customers

When the retailer has to deliver the goods in mobile stores to customers, there are additional trade-offs to consider. The outbound delivery cost typically increases with the distance between stores and customers. As a result, the retailer has the incentive to set smaller influence areas of stores, so that the delivery cost becomes lower. More specifically, we assume the delivery incurs cost d^o per kilometer of distance between a store at x_{tj} and a customer at $x \in \mathcal{X}_{tj}$. The outbound delivery cost for a store serving area $\mathcal{X}_{tj}$ is $\int_{x\in\mathcal{X}_{tj}} b^o d(x_{tj},x)\rho_\theta(x)dx$. The outbound delivery cost density, denoted by φ^o, is thus given by

$$\varphi^o\Big(d(x_{tj},x),\rho_\theta(x);W_t(x)\Big) = b^o d(x_{tj},x)\rho_\theta(x).$$

The related CA function can be evaluated by averaging the distance $d(x_{tj},x)$ using $\frac{2}{3\sqrt{\pi}}\sqrt{z_t(x)}$, i.e.,

$$\varphi^o\Big(\frac{2}{3\sqrt{\pi}}\sqrt{z_t(x)},\rho_\theta(x);W_t(x)\Big) = b^o\frac{2}{3\sqrt{\pi}}\sqrt{z_t(x)}\rho_\theta(x).$$

Hereafter we omit the constant $\frac{2}{3\sqrt{\pi}}$ for brevity since we can include it in the constant b^o. By incorporating the outbound delivery cost density into the CA profit density function, as described in (6) and (7), we have the following result:

$$\begin{aligned} r^{\psi}_{\theta}(z_t;W_t) &= \int_{x\in\mathcal{X}_t}\psi_\theta(z_t(x);W_t(x))dx \\ &= \int_{x\in\mathcal{X}_t}\left(\bar{r}\rho_\theta(x)-\beta_{\mathsf{TSP}}\frac{c_t}{S}\rho_\theta(x)\sqrt{z_t(x)}-\frac{a^f\rho_\theta(x)z_t(x)+b^f}{z_t(x)}-b^o\rho_\theta(x)\sqrt{z_t(x)}\right)dx. \end{aligned}$$

The optimal solution $z^*_t(x;\theta)$ and the optimal profit density can be point-wisely obtained by

$$\begin{aligned} z^*_t(x;\theta) &= \left(\frac{2b^f S}{(\beta_{\mathsf{TSP}}c_t+b^o S)\rho_\theta(x)}\right)^{\frac{2}{3}}, \\ \psi_\theta(z^*_t(x;\theta);W_t(x)) &= (\bar{r}-a^f)\rho_\theta(x)-3\left(b^f\right)^{\frac{1}{3}}\left(\frac{\beta_{\mathsf{TSP}}c_t+b^o S}{2S}\rho_\theta(x)\right)^{\frac{2}{3}}. \end{aligned}$$

We next analyze the effect of outbound delivery on the CA gap. In addition to Assumptions 1–2, we conclude the following technical assumption from the scenario of delivery to customers.

Assumption 6.

(I) The delivery cost φ^o is a convex function of the potential trip distance $d(x_j,x)$.

(II) The second derivative of the delivery cost φ^o with respect to $d(x_j,x)$ exists, and its absolute value is bounded from above by η^o.

(III) Let X be a random location in $\mathcal{X}_j$. The variance of random delivery distance $d(x_j,X)$ is denoted by $Var(d(x_j,X);\mathcal{X}_j)$.

(IV) Within each influence area j, the difference between the average delivery trip distance and $(2/(3\sqrt{\pi}))\sqrt{|\mathcal{X}_j|}$ is bounded within $[0, \eta^d]$.

(V) The delivery cost φ^o is Lipschitz continuous on $d(x_j, x)$ with modulus L_φ.

As to the outbound delivery cost, the CA is to replace the delivery trip distance $d(x_j, x)$ with $2/(3\sqrt{\pi})\sqrt{|\mathcal{X}_j|}$. The latter is the average distance from the center of a round disk with area $|\mathcal{X}_j|$ to a point on this disk. However, in general, the average delivery trip distance, denoted by $\bar{d}_j$, is not necessarily equal to $2/(3\sqrt{\pi})\sqrt{|\mathcal{X}_j|}$ since the influence area is not a circle in practice. If the CA were instead to replace the delivery trip distance $d(x_j, x)$ with $\bar{d}_j$, then we would obtain

$$0 \leq \int_{\mathcal{X}_j} \varphi^o\left(d(x_j, x), \rho(x)\right) dx - \int_{\mathcal{X}_j} \varphi^o(\bar{d}_j, \rho(x)) dx \leq \frac{\eta^o}{2} Var(d(x_j, X); \mathcal{X}_j). \tag{34}$$

Here the first inequality is due to the convexity of φ^o in $d(x_j, x)$ and Jensen's inequality (Assumption 6(I)). The second inequality is due to the sharpened Jensen's inequality and Assumptions 6(II) and (III).

It remains to quantify the error induced by using $\bar{d}_j$ instead of $2/(3\sqrt{\pi})\sqrt{|\mathcal{X}_j|}$ in (34). Following Assumption 6(IV) and (V), this error is bounded as follows:

$$0 \leq \int_{\mathcal{X}_j} \left(\varphi^o(\bar{d}_j, \rho(x)) - \varphi^o\big(\frac{2}{3\sqrt{\pi}}\sqrt{|\mathcal{X}_j|}, \rho(x)\big)\right) dx \leq \int_{\mathcal{X}_j} L_\varphi \eta^d dx = L_\varphi \eta^d |\mathcal{X}_j|. \tag{35}$$

The first inequality is due to $\bar{d}_j \geq 2/(3\sqrt{\pi})\sqrt{|\mathcal{X}_j|}$ and the fact that the outbound cost is non-decreasing in delivery trip distance. Combining Theorem 1 with (34)–(35), we obtain the modified univeral bounds for the CA gap in the scenario of delivery to customers as follows:

$$\beta_{\mathsf{CA}} := \sup_{\{\boldsymbol{x}, \boldsymbol{\mathcal{X}}\}} \sum_j \Bigg(\frac{\eta^\psi}{2} Var(z^*; \mathcal{X}_j) + \frac{\eta^i}{2} Var(d(X', X); \mathcal{X}_j) + \frac{\eta^\rho}{2} Var(\rho(X); \mathcal{X}_j) + \frac{\eta^o}{2} Var(d(x_j, X); \mathcal{X}_j) + L_\varphi \eta^d |\mathcal{X}_j| \Bigg).$$

We now proceed to analyze the regret bound for proposed algorithms in this scenario of delivery to customers. Note that in the mobile retail store location problem involving delivery, the outbound delivery cost exhibits a similarity to Assumption 4(VI). If we denote the Lipschitz constant of φ^o by α_o, Lemma 9 will still hold by letting $L_r = \bar{r} + |\mathcal{X}|\alpha_i + |\mathcal{X}|\alpha_f + |\mathcal{X}|\alpha_o$. Therefore, one can opt to use CA-O Learning. or CA-O Faster Learning to solve this problem, and the regret performance remains bounded by Theorem 2 or Theorem 3, respectively.

E.3 Last-mile delivery with micro-depots

Finally, we investigate a micro-depot location problem for last-mile delivery. The application of micro-depots arises directly from the context of urban logistics. Specifically, we adopt the problem of using crowdsourced ride-share mobility for last-mile package deliveries, as developed in Qi et al. (2018). The decision involves delimiting the entire service region into individual zones, with a micro-depot centered at each zone as the trans-shipment terminal. A truck loads packages from a distribution center at x_d and traverses terminals to unload packages. Within each zone, idle ride-share vehicles are paid to pick up packages at the micro-depot and fulfill the last-mile deliveries. The objective is to minimize costs incurred. The problem involves intricate modeling of the open-loop vehicle route lengths and the driver compensation schemes. However, the CA model ultimately can be reduced to the following form:

$$\min_{z_t(\cdot)} \int_{x\in\mathcal{X}_t} \psi_\theta(z_t(x); W_t(x))dx$$
$$= \int_{x\in\mathcal{X}_t} \left(\varphi^i\Big(d(x_d,x), \rho_\theta(x) z_t(x); W_t(x)\Big) + \varphi^o\Big(\frac{\sqrt{2}}{3}\sqrt{z_t(x)}, \rho_\theta(x); W_t(x)\Big) \right) dx,$$

in which the inbound trucking cost and the outbound delivery cost are given by

$$\varphi^i\Big(d(x_d,x), \rho_\theta(x) z_t(x); W_t(x)\Big) = \frac{U}{\sqrt{z_t(x)}} + V d(x_d,x)\rho_\theta(x),$$
$$\varphi^o\Big(\frac{\sqrt{2}}{3}\sqrt{z_t(x)}, \rho_\theta(x); W_t(x)\Big) = \big(E\sqrt{\rho_\theta(x)} + F\big)\rho_\theta(x)\sqrt{z_t(x)} + G\rho_\theta(x) + H\sqrt{\rho_\theta(x)},$$

respectively, where U, V, E, F, G, and H are constant numbers composed of system parameters. (See Equations (11), (12), (9), and (10) in Qi et al. (2018) for detailed derivations.) We also adopt the calibrated parameter values from Qi et al. (2018): $U = 0.9580$, $V = 0.0045$, $E = 5.1167 \times 10^{-4}$, $F = 0.0250$, $G = 0.1422$, and $H = 1.4092$). Determining the value of these constants involves approximating vehicle route lengths. We ignore this approximation error, given not just the scope of this paper but also the practice in the literature of location-routing problems. Also note that, in this problem, we follow the assumption in Qi et al. (2018) of using "Manhattan distance"; as a result, the mean distance from the center of a zone (which is now a rhombus) with size $z_t(x)$ to a random point in the zone is $\sqrt{2}\sqrt{z_t(x)}/3$ instead of $2/(3\sqrt{\pi})\sqrt{z_t(x)}$. Similar to the previous application, applying the first-order condition yields the optimal solution and the optimal cost density function, as follows:

$$z_t^*(x;\theta) = \frac{U}{\big(E\sqrt{\rho_\theta(x)} + F\big)\rho_\theta(x)},$$
$$\psi_\theta^*(z_t^*(x;\theta); W_t(x)) = 2\left(UE\left(\rho_\theta(x)\right)^{\frac{3}{2}} + UF\rho_\theta(x)\right)^{\frac{1}{2}} + Vd(0,x)\rho_\theta(x) + G\rho_\theta(x) + H(\rho_\theta(x))^{\frac{1}{2}}.$$

We solve an online version of this problem, in which the decisions of micro-depot siting or, equivalently, service zoning are adjusted over time t. We run CA-O Faster Learning 500 times. Figure 7 shows the evolution of the regret. The blue shaded area is the 95% confidence interval. The sublinear trend of the regret is consistent with the theory given by Theorem 3 without the linear term $2\beta_{\mathsf{CA}}T$. (Here, the CA gap is irrelevant because we omit the design dicretization and directly use the objective function of the CA model.) In addition, the optimization problem involved in the learning process can be solved within seconds by CA-O Faster Learning at each round. To sum up, the sublinear regret and the high computational efficiency suggest that our proposed framework potentially is widely applicable in solving practical online facility location problems.

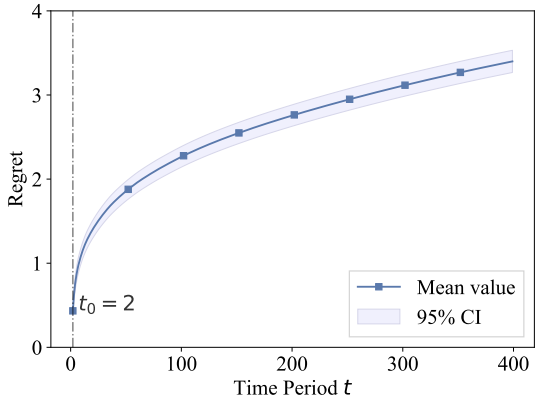


Figure 7: Regret of cost in the online micro-depot location problem.

References

Rajeev Agrawal. The continuum-armed bandit problem. SIAM journal on control and optimization, 33(6): 1926–1951, 1995.

Shipra Agrawal and Nikhil R Devanur. Bandits with global convex constraints and objective. Operations Research, 67(5):1486–1502, 2019.

Sina Ansari, Mehmet Başdere, Xiaopeng Li, Yanfeng Ouyang, and Karen Smilowitz. Advancements in continuous approximation models for logistics and transportation systems: 1996–2016. Transportation Research Part B: Methodological, 107:229–252, 2018.

D Applegate, W Cook, DS Johnson, and NJA Sloane. Using large-scale computation to estimate the beardwood-halton-hammersley tsp constant. Presentation at Symposium on Operations Research, 42 SBPO, 2010.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. Operations Research, 67(1):90–108, 2019.

Jillian Beardwood, John H Halton, and John Michael Hammersley. The shortest path through many points. Mathematical Proceedings of the Cambridge Philosophical Society, 55(4):299–327, 1959.

Elena Belavina. Grocery store density and food waste. Manufacturing & Service Operations Management, 23 (1):1–18, 2021.

Oded Berman, Dimitris Bertsimas, and Richard C Larson. Locating discretionary service facilities, ii: maximizing market size, minimizing inconvenience. Operations Research, 43(4):623–632, 1995.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. Management Science, 66(3): 1025–1044, 2020.

Shahzad F Bhatti, Michael K Lim, and Ho-Yin Mak. Alternative fuel station location model with demand learning. Annals of Operations Research, 230(1):105–127, 2015.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. The Journal of Machine Learning Research, 22(1):5928–5976, 2021.

Moïse Blanchard, Alexandre Jacquillat, and Patrick Jaillet. Probabilistic bounds on the k-traveling salesman problem and the traveling repairman problem. Mathematics of Operations Research, 49(2):1169–1191, 2024.

Bloomberg. Robo-Vans could start delivering profits and not just Pizza. https://www.bloomberg.com/news/newsletters/2021-08-09/robo-vans-could-start-delivering-profits-and-not-just-pizza, August 2021. [Bloomberg News; Online; accessed on 9-15-2023].

Alireza Boloori and Reza Zanjirani Farahani. Facility location dynamics: An overview of classifications and applications. Computers & Industrial Engineering, 62(1):408–420, 02 2012.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. Journal of Machine Learning Research, 12(5):1655–1695, 2011.

Cem Canel, Basheer M. Khumawala, Japhett Law, and Anthony Loh. An algorithm for the capacitated, multi-commodity multi-period facility location problem. Computers & Operations Research, 28(5):411–427, 2001.

Junyu Cao and Wei Qi. Stall economy: The value of mobility in retail on wheels. Operations Research, 71(2): 708–726, 2023.

John Gunnar Carlsson and Siyuan Song. Coordinated logistics with a truck and a drone. Management Science, 64(9):4052–4069, 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 208–214. PMLR, 2011.

Alon Cohen, Tamir Hazan, and Tomer Koren. Tight bounds for bandit combinatorial optimization. In Proceedings of the 2017 Conference on Learning Theory, pages 629–642. PMLR, 2017.

Carlos F Daganzo. Logistics Systems Analysis. Springer Science & Business Media, 2005.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In 21st Annual Conference on Learning Theory, pages 355–366, 2008.

DPD 2023. New microdepot: DPD delivers by cargo bike in Dresden. https://www.dpd.com/de/en/news/new-microdepot-dpd-delivers-by-cargo-bike-in-dresden/, 2023. [DPD; Online; accessed on 11-28-2023].

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". Management Science, 68(1):9–26, 2022.

Forbes. Stores on wheels startup Robomart gains momentum with Unilever partnership. https://www.forbes.com/sites/joanverdon/2022/06/02/stores-on-wheels-startup-robomart-gains-momentum-with-unilever-partnership/, 2022. [Forbes Magazine; Online; accessed on 6-30-2023].

FT. SoftBank invests $940m in driverless delivery start-up Nuro. https://www.ft.com/content/4f06e7ca-2e0c-11e9-8744-e7016697f225, February 2019. [Financial Times; Online; accessed on 7-01-2023].

Chloe Kim Glaeser, Marshall Fisher, and Xuanming Su. Optimal retail location: Empirical methodology and application to practice. Manufacturing & Service Operations Management, 21(1):86–102, 2019.

Xiangyu Guo, Janardhan Kulkarni, Shi Li, and Jiayi Xian. On the facility location problem in online and dynamic models. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020), volume 176, pages 42:1–42:23. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Jinhui Han, Ming Hu, and Guohao Shen. Deep neural newsvendor. Available at SSRN 4582188, 2023.

Nam Ho-Nguyen and Fatma Kılınç-Karzan. Risk guarantees for end-to-end prediction and optimization processes. Management Science, 68(12):8680–8698, 2022.

IDTechEx 2020. Mobile EV chargers on the go: Niche or disruptive? https://www.idtechex.com/en/research-article/mobile-ev-chargers-on-the-go-niche-or-disruptive/20533, 2020. [IDTechEx; Online; accessed on 3-19-2021].

Wang Kai, Alexandre Jacquillat, and Vikrant Vaze. Vertiport planning for urban aerial mobility: an adaptive discretization approach. Manufacturing & Service Operations Management, 24(6):3215–3235, 2022.

Olav Kallenberg. Foundations of modern probability, volume 2. Springer, 1997.

Haim Kaplan, David Naori, and Danny Raz. Almost tight bounds for online facility location in the random-order model. In Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1523–1544. SIAM, 2023.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. The Journal of Machine Learning Research, 21 (1):5402–5446, 2020.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In International Conference on Artificial Intelligence and Statistics, pages 2066–2076. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Proceedings of the 34th International Conference on Machine Learning, volume 70, pages 2071–2080. JMLR, 2017.

J. G. Liao and Arthur Berg. Sharpening jensen's inequality. The American Statistician, 73(3):278–281, 2019.

Michael K Lim, Ho-Yin Mak, and Zuo-Jun Max Shen. Agility and proximity considerations in supply chain design. Management Science, 63(4):1026–1041, 2017.

Guodong Lyu and Chung-Piaw Teo. Last mile innovation: The case of the locker alliance network. Manufacturing & Service Operations Management, 24(5):2425–2443, 2022.

Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. IEEE Transactions on Automatic Control, 54(12):2787–2802, 2009.

Adam Meyerson. Online facility location. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pages 426–431. IEEE, 2001.

Sajad Modaresi, Denis Sauré, and Juan Pablo Vielma. Learning in combinatorial optimization: What and how to explore. Operations Research, 68(5):1585–1604, 2020.

Neolix. The infinite Possibilities between Autonomous Vehicle and X. https://www.neolix.net/application, 2023. [Online; accessed on 9-15-2023].

Nuro. Nuro's next-generation vehicle: Customizable compartments. https://www.nuro.ai/vehicle, 2023. [Online; accessed on 7-02-2023].

Yanfeng Ouyang and Carlos F Daganzo. Discretization and validation of the continuum approximation scheme for terminal system design. Transportation Science, 40(1):89–98, 2006.

Wei Qi, Lefei Li, Sheng Liu, and Zuo-Jun Max Shen. Shared mobility for last-mile delivery: Design, operational prescriptions, and environmental impact. Manufacturing & Service Operations Management, 20(4):737–751, 2018.

Robomart. Robomart Models. https://robomart.ai/models, 2023. [Online; accessed on 7-02-2023].

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. Mathematics of Operations Research, 35(2):395–411, 2010.

Ilya O Ryzhov and Warren Powell. The knowledge gradient algorithm for online subset selection. In 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, pages 137–144. IEEE, 2009.

Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. Operations Research, 60(1):180–195, 2012.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. Foundations of computational mathematics, 12(4):389–434, 2012.

Mengxin Wang, Meng Qi, Junyu Cao, and Zuo-Jun Max Shen. Urban courier: Operational innovation and data-driven coverage-and-pricing. Available at SSRN 3678317, 2020.

Xin Wang, Michael K. Lim, and Yanfeng Ouyang. A continuum approximation approach to the dynamic facility location problem in a growing market. Transportation Science, 51(1):343–357, 2017.

George O Wesolowsky. Dynamic facility location. Management Science, 19(11):1241–1248, 1973.

WSJ. Kroger plans to introduce driverless grocery deliveries. https://www.wsj.com/articles/kroger-plans-to-introduce-driverless-grocery-deliveries-1530190801, June 2018. [The Wall Street Journal; Online; accessed on 6-20-2023].

Dennis J. Zhang, Hengchen Dai, Lingxiu Dong, Qian Wu, Lifan Guo, and Xiaofei Liu. The value of pop-up stores on retailing platforms: Evidence from a field experiment with alibaba. Management Science, 65 (11):5142–5151, 2019.