Online Resource Allocation with Non-Stationary Customers

Xiaoyue Zhang [* 1] **Hanzhang Qin** [* 1] **Mabel C. Chou** [1]

Abstract

We propose a novel algorithm for online resource allocation with non-stationary customer arrivals and unknown click-through rates. We assume multiple types of customers arriving in a nonstationary stochastic fashion, with unknown arrival rates in each period. Additionally, customers' click-through rates are assumed to be unknown and only learnable online. By leveraging results from the stochastic contextual bandit with knapsack and online matching with adversarial arrivals, we develop an online scheme to allocate the resources to nonstationary customers. We prove that under mild conditions, our scheme achieves a "best-of-both-world" result: the scheme has a sublinear regret when the customer arrivals are near-stationary, and enjoys an optimal competitive ratio under general (non-stationary) customer arrival distributions. Finally, we conduct extensive numerical experiments to show our approach generates near-optimal revenues for all different customer scenarios.

1. Introduction

The realm of online resource allocation, critical in fields ranging from online advertising to traffic management, poses a substantial challenge: how to effectively and dynamically distribute limited resources in response to ever-evolving consumer behaviors. This task is particularly arduous in environments where consumer preferences fluctuate rapidly, rendering traditional static allocation models ineffective.

Addressing this challenge, we introduce the Unified Learning-while-Earning (ULwE) algorithm, a novel approach embedded within the Contextual Bandit with Knapsacks (CBwK) framework. The ULwE algorithm stands out for its real-time adaptability to changing consumer preferences and uncertain click-through rates, a notable departure from traditional methods that often fail to capture the non-stationary nature of customer arrivals.

This paper is structured as follows: Section 1.1 summarizes our main contributions. Section 1.2 reviews relevant literature, setting the stage for our innovation. Section 2 introduces our models for online resource allocation, adapted for non-stationary environments. Section 3 discusses the ULwE algorithm in detail, presenting our unified theoretical framework. Finally, Section 4 validates our algorithm through empirical studies on simulated data and Section 5 concludes the paper.

1.1. Basic Setup and Contributions

1.1.1. Basic setup

Our online resource allocation problem encompasses n resources, each defined by a revenue parameter (r_i), capacity (c_i), and a latent variable (θ_i) within a predefined space Θ. It operates over T time periods, with each period $t \in [T]$ witnessing the arrival of a customer, characterized by a feature vector $x^t \in \mathbb{R}^d$. The likelihood of a customer purchasing from resource i is modeled by $f_i(x^t, \theta_i^*)$, where θ_i^* is the true value of θ_i, and $f_i(x^t, \theta_i) \leq 1$ ensures model validity. We assume the θ_i^*s are unknown a priori, and we use $|\Theta|$ to denote the cardinality of the parameter space Θ.

The context vector x^t for each time period t is independently drawn from an unknown distribution $P(x^t = x^{(l)}) = \mu_l^t$ for each customer type $l \in 1, 2, \ldots, L$. Given the variability of this distribution across different time periods, the customer arrival process is inherently non-stationary. For every customer type represented by $x^{(l)}$, the expected number of arrivals, denoted as λ_l, is computed over the time horizon as $\lambda_l = \sum_{t=1}^{T} \mu_l^t$. We assume all μ_l^ts are unknown but the λ_ls are given in the beginning. In practice, λ_l can be derived from historical data using machine learning techniques and can usually be estimated with high accuracy. For instance, a regression algorithm was employed on prediction traffic data for online advertising, resulting in a total prediction error rate of approximately 0.9% (Lai et al., 2016). This information also can be regarded as advice, and it is

*Equal contribution. Author ordering determined by coin flip. [1]Institute of Operations Research and Analytics, National University of Singapore, Singapore 117602. Correspondence to: Hanzhang Qin <hzqin@nus.edu.sg>, Mabel C. Chou <mabelchou@nus.edu.sg>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

shown by Lyu & Cheung (2023) without any advice the nonstationary BwK (CBwK with identical contexts) admits a worst-case regret linear in T. We will discuss the scenario where the knowledge about λ_l is approximately true in the Appendix E.

1.1.2. Main Contributions

Our primary contribution lies in the development of the ULwE algorithm, a novel approach within the CBwK framework under non-stationary arrivals. This algorithm uniquely offers simultaneous guarantees of sub-linear regret and a constant competitive ratio (CR), effectively addressing the complexities of unknown click-through rates and non-stationary customer arrivals in dynamic environments.

A Unified Learning-while-Earning Algorithm Framework with Sublinear Regret and Constant CR. We propose a unified study of two essential performance guarantees in online resource allocation: regret and CR. Achieving sublinear regret is crucial in online advertising and resource allocation systems, as it signifies that the regret incurred by the algorithm grows at a slower rate compared to the time horizon. Additionally, the constant CR ensures that our algorithm achieves a consistently high-performance level in comparison to an optimal offline policy. While each guarantee has been explored separately, our main contribution lies in investigating them together within a unified algorithm framework and achieving the best of both worlds.

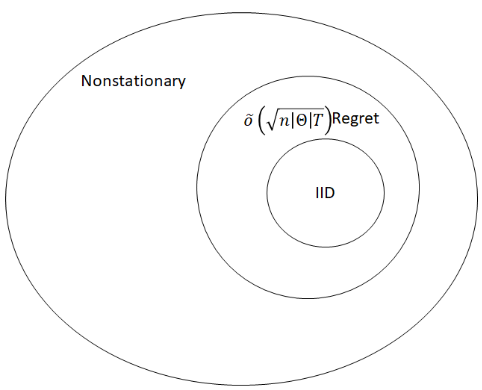

Figure 1. Problem categories

Figure 1 illustrates a graphical representation of three ellipsoids to visually depict this concept. The smallest ellipsoid represents the problem category of i.i.d. arrivals. The intermediate ellipsoid represents problems under near-i.i.d. arrivals scenario, in which our algorithm achieves an expected regret bound of $\tilde{O}(\sqrt{n|\Theta|T})$. The largest ellipsoid encompasses all types of nonstationary customer arrivals, including the most challenging scenarios. In this case, our algorithm provides a unified regret bound, which offers sublinear regret with a constant CR performance, allowing for effective resource allocation even in highly dynamic environments. Overall, under nonstationary arrivals, our ULwE algorithm recovers $\tilde{O}(\sqrt{n|\Theta|T})$ expected regret under near-i.i.d. arrivals (the exact condition is specified in Section 3) and guarantees in general

$$\text{OPT} \leq \left(1 + \frac{(1+\min_{i\in[n]} c_i)\left(1-e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\right)\mathbb{E}[\text{ALG}] + \tilde{O}(\sqrt{n|\Theta|T}),$$

where OPT denotes the expected revenue the optimal online algorithm that knows the true latent variable achieves, ALG denotes the revenue our proposed online learning algorithm achieves (the latent variable is unknown in priori). When $\min_{i\in[n]} c_i$ is sufficiently large, it can be shown the above guarantee can be rewritten as $\text{OPT} \leq (1+1/(1-1/e))\mathbb{E}[\text{ALG}] + \tilde{O}(\sqrt{n|\Theta|T})$, or equivalently, $\mathbb{E}[\text{ALG}] \geq (1-1/(2-1/e))\text{OPT} - \tilde{O}(\sqrt{n|\Theta|T})$. This guarantee is the tightest possible given the underlying contextual arrival process is adversarial and the click-through rates are unknown (see Theorem 4, Cheung et al. 2022).

Non-Stationary Customer Arrivals. Our approach to non-stationary customer arrivals encompasses two distinct scenarios: those with stochastic distributions and adversarial ones. For stochastic distributions, we employ past data to create empirical distributions, aiding in future predictions. This is achieved using the Upper Confidence Bound (UCB) algorithm and the Deterministic Linear Programming (DLP) formulation, optimizing resource allocation by maximizing expected revenue within the confidence bounds of demand rates.

In contrast, for adversarial or unknown distribution scenarios, we adopt a greedy algorithm, as suggested by Cheung et al. (2022), which factors in discounted revenue and inventory constraints. This method ensures efficient allocation by considering both revenue maximization and resource limitations.

Our ULwE algorithm amalgamates these strategies, seamlessly transitioning between them based on the nature of customer arrivals. This integration, alongside inbuilt condition checks, endows our algorithm with the versatility to operate effectively across varying arrival patterns without prior knowledge of their sequence types. Such adaptability makes it well-suited for dynamic real-world contexts, addressing the challenges posed by diverse customer arrival processes.

Unknown Click-Through Rates. Our algorithm addresses the complexities introduced by unknown click-through rates in online advertising and resource allocation systems. It dynamically estimates these rates by leveraging historical data alongside current information, ensuring adaptive and informed decision-making.

Central to our approach is the use of a parametric model with a pre-defined prior distribution for the parameter $\theta \in \Theta$, allowing us to capture the nuanced relationships between different customer types and their purchasing probabilities. Initially, our algorithm samples a subset of θ values from existing data to construct the set Θ, which is then continuously refined as new data becomes available. This process enables our algorithm to evolve with and respond to the changing click-through rate landscape, ensuring resource allocation decisions are both current and data-informed.

This strategy of integrating historical insights with real-time updates provides a well-rounded perspective on customer behavior, crucial for optimizing resource allocation in the dynamic field of online advertising.

1.2. Literature Review

Contextual Bandit with Knapsack under Non-Stationary Arrivals. CBwK problems in non-stationary environments, integral to our research, have been a focal point in recent studies. Chen et al. (2013) were pioneers in CBwK within online advertising, proposing a novel greedy algorithm focusing on the reward-to-cost ratio. Badanidiyuru et al. (2014) introduced a probabilistic method, selecting optimal strategies from policy sets. Agrawal et al. (2015) followed with a computationally efficient variant. Complementing these, Agrawal & Devanur (2015) developed LinCBwK, which utilizes optimistic estimates from confidence ellipsoids to adjust rewards. Recently, Slivkins & Foster (2022) integrated LagrangeBwK (Immorlica et al., 2019) with SquareCB, merging computational efficiency with statistical optimality.

In tackling non-stationary problems, Wei & Luo (2018) proposed an algorithm that achieves $\sqrt{T}$ regret in the adversarial setting and $log(T)$ regret in the stochastic setting. This result was further advanced to optimal by Zimmert & Seldin (2019; 2021). Moreover, Jin et al. (2023) achieved similar results without relying on certain assumptions. Luo et al. (2023) investigated high-probability regret bounds in adversarial settings and applied it to contextual bandits. The adversarial version of BwK, explored by Immorlica et al. (2019). Unlike scenarios with diminishing regret, these algorithms are bound to achieve approximate ratios (Kesselheim & Singla, 2020; Fikioris & Tardos, 2023), even in environments with a single switch (Immorlica et al., 2022). The understanding of the approximation-ratio aspect is now well-established (Castiglioni et al., 2022; Sivakumar et al., 2022; Liu et al., 2022).

Online Resource Allocation. Our work intersects significantly with online resource allocation (also known as the AdWord problem, see Mehta et al. (2007)), particularly in the assignment of resources to real-time, random demands. The CR serves as a key performance metric in this domain. Karp et al. (1990) investigated one-sided bipartite arrivals revealed the optimality of a simple $1/2-$competitive greedy algorithm among deterministic approaches. On the Adwords problem, Mehta et al. (2005) modeled as a linear program (LP), led to an online algorithm with a $1 - 1/e$ CR based on primal-dual LP analysis. Fahrbach et al. (2022) further broke new ground with the online correlated selection subroutine reaching a CR of at least 0.5086.

The works most aligned with ours are those by Ferreira et al. (2018); Zhalechian et al. (2022); Cheung et al. (2022). Ferreira et al. (2018) proposed a dynamic algorithms highlighted versatility in handling resource-constrained bandit problems. Zhalechian et al. (2022) addressed personalized learning resource allocation with a Bayesian regret analysis, adept for adversarial customer contexts. Cheung et al. (2022) focused on the allocation of limited resources over time to heterogeneous customers, a methodology that has significantly influenced our approach to CR analysis.

2. Model

Our objective is to maximize the total reward obtained from customers' clicks or purchases while considering budget constraints for the available resources or advertisers. We consider a system with n resources, resource $i \in [n]$ has an initial budget value c_i. Customers arrive at the system from period 1 to T. When a customer arrives, we observe the feature vector $x \in \mathbb{R}^d$ of the customer, where d is the length of the feature vector. At this point, the system must make a decision: either reject the customer irrevocably or assign a resource to the customer immediately. If the resource j is assigned, the customer's likelihood of purchasing the resource is determined by the probability function $f_j(x)$ associated with that resource j and the customer's feature vector x. If the customer makes a purchase, the system earns a reward r_j that is deducted from the budget of the assigned resource. If the customer does not make a purchase, no reward is earned.

To address this complex problem, our model incorporates two key elements: non-stationary customer arrivals and unknown click-through rates. We recognize that customer behavior may change over time, requiring the system to adapt to these fluctuations. Additionally, the exact click-through rates, which represent the likelihood of customers clicking on a resource or making a purchase, are initially unknown. However, we utilize extensive data to estimate and update these rates in real-time.

2.1. Handling Non-Stationary Customer Arrivals

In real-world systems, customer arrivals often exhibit non-stationary behavior, reflecting temporal variations in customer types. For example, students accessing a system are likely to show different patterns during mornings, evenings, and class hours. We categorize non-stationary customer arrivals into two types for our model:

- **Stochastic Arrivals with Non-Stationary Distributions:** These are regular scenarios without disruptive events. Past data can be used to learn distributions for future arrivals.
- **Adversarial Arrivals with Agnostic Distributions:** Such scenarios occur during unpredictable events (e.g., Black Friday), where arrivals are assumed to be controlled by an adversary.

We assume no prior knowledge of arrival patterns except for the total arrival rate λ_j for each customer type $j \in [L]$. This approach caters to both i.i.d. and adversarial sequences, overcoming the limitations of prior models which often cannot predict the nature of future arrivals.

2.2. Modeling Unknown Click-Through Rates

Addressing unknown click-through rates (CTR) is pivotal, particularly in contexts with ample historical data. While historical data provides insights into individual preferences, actual CTRs are influenced by recent factors like ad quality and competitor promotions. Our model adopts a parametric approach, beginning with a known prior distribution for $\theta \in \Theta$. This setup allows for a flexible, parametric formulation of the purchase probability function $f(x^t, \theta)$, where $x^t \in \mathbb{R}^d$ is the feature vector for customer arrivals at time $t \in [T]$.

An example model choice is the polynomial logistic regression model (Richardson et al., 2007), with θ representing polynomial term parameters. To utilize historical data, our algorithm samples a subset of θ values initially, forming the set Θ. This set is dynamically updated over time, enabling the algorithm to adapt to changing conditions and refine decision-making processes continually. In practice, Θ can include any type of the parameters involved in a parameter set associated with a particular machine learning model (e.g., a Large Language Model) for CTR prediction.

2.3. An Upper Bound on the Optimal Revenue

In this section, our primary objective is to establish a theoretical upper bound on the optimal revenue in our model, denoted as $\overline{\text{OPT}}$. This upper bound provides a crucial benchmark for assessing the performance of various algorithms and understanding their efficiency and effectiveness in different resource allocation scenarios.

To establish an upper bound, we introduce a deterministic linear programming model J^D, which aims to maximize the deterministic revenue given the capacities and customer arrival patterns:

$$\begin{aligned} J^D(c,t) = &\max_{s_{ij}, i\in[n], j\in[L]} \sum_{i=1}^{n}\sum_{j=1}^{L}\sum_{\tau=1}^{t} r_i s_{ij}^{\tau} f_i(x^j,\theta^*) \\ &\text{s.t. } \sum_{j=1}^{L}\sum_{\tau=1}^{t} {\mu_j}^{\tau} s_{ij}^{\tau} f_i(x^j,\theta^*) \le c_i, \forall i \in [n] \\ &\sum_{i=1}^{n}\sum_{\tau=1}^{t} s_{ij}^{\tau} = \sum_{\tau=1}^{t} {\mu_j}^{\tau}, \forall j \in [L] \\ &s_{ij}^{\tau} \ge 0, \forall i \in [n], \forall j \in [L] \end{aligned} \tag{1}$$

We establish that:

$$\text{OPT} \le J^D(c,t),$$

indicating that $J^D(c,t)$ serves as an upper bound for the optimal revenue. Consequently, we define regret as the difference between this upper bound and the actual optimal revenue achieved, formulated as: Regret $= \overline{\text{OPT}} - \text{ALG}$.

This quantifies the efficiency of different algorithms in approaching the theoretical maximum revenue. For a comprehensive understanding of the problem formulation, constraints, and detailed derivations, please refer to Appendix B, which includes the full mathematical treatment and theoretical analysis supporting these findings.

3. Algorithm and Analysis

We propose the ULwE algorithm, displayed in Algorithm 1, which is applicable in both stochastic and adversarial environments. The algorithm design involve constructing switch strategy to address the model uncertainty on μ, as discussed in Section 3.1. In Section 3.2, we provide a regret upper bound to ULwE, and demonstrate the scheme has a sublinear regret when the customer arrivals are near-stationary, and enjoys an optimal competitive ratio under general (non-stationary) customer arrival distributions. In Section 3.3 we provide a sketch proof of the regret upper bound, where the complete proof is in Appendix D

3.1. The ULwE Algorithm

In this section, we explore the ULwE algorithm (Algorithm 1), which is applicable in both stochastic and adversarial environments. ULwE is an evolution of the UCB paradigm, integrating aspects of inventory balancing with a penalty function (Golrezaei et al., 2014). A distinctive feature of ULwE is its dynamic approach: the algorithm continually updates the parameter set θ and monitors capacity constraints to determine the necessity of a policy switch.

Algorithm 1 Unified Learning-while-Earning (ULwE)

Input: Resource capacities c, time horizon T, customer types L
Initialize $\Omega_i^0 = \Theta_i$ for all $i \in [n]$
for $t = 1$ **to** T **do**
  if switch = FALSE **then**
    $I^t = \text{ALG}_{\text{LP}}(c, T, L, \Omega_i^{t-1})$.
    Check if conditions (5) and (6) for switching are met.
    if any condition is violated **then**
      switch = TRUE.
    end if
  else
    $I^t = \text{ALG}_{\text{ADV}}(c, T, L, \Omega_i^{t-1})$.
  end if
  Check if conditions (3) and (4) for updating Ω_i are met.
  if any condition is violated **then**
    Remove $\bar{\theta}^t$ from $\Omega_{I^t}^t$
  else
    Set $\Omega_i^t = \Omega_i^{t-1}$ for all $i \in [n]$
  end if
end for

Specifically, ULwE adapts its policy to best suit customer patterns, whether they follow i.i.d. or adversarial arrival sequences. This strategy of switching sets ULwE apart from the traditional UCB-based methods used in the stochastic context (Agrawal & Devanur, 2015) and the strategies applied to the adversarial BwK settings (Immorlica et al., 2019). While ULwE shares similarities with previous works on stochastic and adversarial bandits with knapsacks in the methodologies of arm selection, which involves LP optimization and inventory balancing, and in its fundamental updating mechanism, which is based on the UCB approach.

The algorithm operates through three crucial stages: selecting an arm to obtain rewards and consumption vectors, updating the confidence set Θ, and assessing whether to switch states.

Arm Selection. In each period t, our algorithm offers a resource I^t from a set of n resources to the customer. This process constitutes the arm selection phase, which is executed differently under two distinct protocols: ALG_{LP} and ALG_{ADV}, catering to different customer arrival patterns.

For ALG_{LP} protocol (Algorithm 2) designed for environments with stationary customer arrivals, we assume that the probability of each customer type l arriving at time t remains constant i.e. $\mu_l^t = \mu_l$ for all $t \in [T]$. The algorithm solves a LP problem at each step to maximize revenue, constrained by inventory limits and the requirement that the sum of probabilities of offering all resources to each customer

Algorithm 2 ALG_{LP} Protocol

Input: Resource capacities c, time horizon T, customer types L
Initialize $\Omega_i^0 = \Theta_i$ for all $i \in [n]$
for $t = 1$ **to** T **do**
  Solve LP in Equation (2) to obtain $\bar{s}^t, \bar{\gamma}^t$
  Select resource i with probability $\bar{s}_{iJ^t}$.
end for

Algorithm 3 ALG_{ADV} Protocol

Input: Resource capacities c, time horizon T, customer types L
Initialize $\Omega_i^0 = \Theta_i$ for all $i \in [n]$
for $t = 1$ **to** T **do**
  Observe the context x^t of the new arrival in period t.
  Select $I^t = \arg\max_{i \in [n]} r_i^t \bar{f}_i(x^t, \Omega_i^{t-1})$.
end for

equals one. The LP formulation is as follows: $U^t =:$

$$\begin{aligned}\max_{s_{ij}, i \in [n], j \in [L]} & \sum_{i \in [n]} r_i \sum_{j \in [L]} \lambda_j s_{ij} \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \\ \text{s.t.} & \sum_{j \in [L]} \lambda_j s_{ij} \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \leq c_i, \ \forall i \in [n] \\ & \sum_{i \in [n]} s_{ij} = 1, \ \forall j \in [L] \\ & s_{ij} \geq 0, \ \forall i \in [n], j \in [L].\end{aligned} \tag{2}$$

Here, the function $\bar{f}_i(x, \Omega)$ calculates the maximum purchase probability for resource i given a set Ω of valid latent variables, i.e. $\bar{f}_i(x, \Omega) = \max_{w \in \Omega}\{f_i(x, w)\}$.

In contrast, ALG_{ADV} protocol (Algorithm 3) is designed to handle the dynamic and unpredictable nature of customer behavior and preferences. In this situation, the underlying distribution of x_t no longer exists. Instead, the value of x_t is chosen by an adversary in each period t. Algorithm 3 computes a real-valued function, $\Psi(x) := \frac{e^x - 1}{e - 1}$, defined over the interval $[0, 1]$. This function is instrumental in adjusting the revenue values r_i^t for each resource i, taking into account the resource's past utilization and its capacity. The adjusted revenue, denoted by r_i^t, is calculated as $r_i \times \left(1 - \Psi\left(\frac{N_i^{t-1}}{c_i}\right)\right)$, where N_i^{t-1} represents the previous consumption of resource i. The algorithm then selects the resource I^t that maximizes the upper confidence bound of the single period revenue, taking into account the modified revenue values and the confidence intervals defined by the set of valid latent variables Ω_i^{t-1}.

Update Confidence Set. In the concept of the UCB strategy, a key aspect is the continuous updating of its confidence set. In our algorithm, this process begins with the formation of a theta set Θ, encompassing a set of possible parameter values. From this set, the confidence set Ω_i^t is constructed. After selecting the resource, the algorithm identifies the maximizer $\bar{\theta}$ of the purchase probability for that resource. The updating mechanism involves two critical equations, for all $\theta \in \Omega^{t-1}$:

$$\left| \sum_{t' \in D_{I^t}^t(\bar{\theta}^t)} (f_{I^{t'}}(x^{t'}, \bar{\theta}^t) - a^{t'}) \right| \leq \sqrt{t \log \frac{2t}{\beta(n,T)}} \tag{3}$$

$$\left| \sum_{t' \in D_{I^t}^t(\bar{\theta}^t)} (f_{I^{t'}}(x^{t'}, \bar{\theta}^t) - f_{I^{t'}}(x^{t'}, \theta)) \right| \leq \sqrt{t \log \frac{2t}{\beta(n,T)}} \tag{4}$$

Here, $\beta(n,T)$ is set as $1/(nT)$. The set D_i^t tracks the periods up to time t where resource i is offered. The customer's purchase decision is represented by $a^t \in \{0,1\}$.

Equation (3) ensures that the accumulated regret, the difference between predicted and actual purchases, remains within a predefined threshold. If this threshold is exceeded, $\bar{\theta}$ is removed from the confidence set, refining the purchase probability estimates. Equation (4) is crucial in maintaining the stability of the algorithm under near-i.i.d. arrival patterns. It limits the occurrence of a switch, fostering a more consistent and effective resource allocation strategy and leading to sublinear regret in these situations.

At its core, this iterative refinement process sharpens the focus of the confidence interval on the most promising parameter values by discarding less probable theta values. This refined approach, supported by continuously updated data, significantly enhances the algorithm's ability to make efficient and accurate resource allocation decisions.

In Section 1.1.1, we employ θ_i^* as a unique latent variable for each resource in the formulation $f_i(x^t, \theta_i^*)$. To simplify the analysis in subsequent sections, we transition to a model that employs a collective latent variable affecting all resources, enhancing analytical tractability.

Switch State Checking. The ULwE algorithm checks if it has switched or not in each time step. When not switched, the algorithm checks two conditions. The switching conditions involve two key inequalities which are checked at each time step t. For all $\theta \in \Omega^{t-1}$:

$$\left| \sum_{l=1}^{t} \sum_{i=1}^{n} r_i (s_{iJ^l}(\theta) - \bar{s}_{iJ^t}^l) f_{iJ^l}(x^{(J^t)}, \theta) \right| \leq \max_{i \in [n]} r_i \sqrt{32t \log(4|\Theta|t/\beta(n,T))}, \tag{5}$$

For all $i \in [n]$

$$\sum_{l=1}^{t} \bar{s}_{iJ^l}^l \bar{f}_i(x^{(J^l)}, \Omega^{l-1}) \leq \frac{t}{T} c_i + \sqrt{2t \log(2t/\beta(n,T))} \tag{6}$$

Equation (5) ensures that the estimated upper bound regret does not grow beyond a sublinear rate. This condition measures the algorithm's performance in terms of regret, which quantifies the deviation between the total expected reward obtained by the algorithm and the maximum achievable reward that an optimal offline algorithm could obtain. Equation (6) checks whether the remaining capacity of the resources is sufficient. This condition ensures that the cumulative resource allocation does not exceed the available capacity plus a certain tolerance level. It takes into account the capacity constraints imposed by the resources and ensures that the algorithm does not allocate more resources than what is available. If both conditions hold, the algorithm continues to the next time step. However, if any of the conditions is violated, indicating either excessive regret growth or insufficient resource capacity, the algorithm switches from the current algorithm (ALG_{LP}) to ALG_{ADV}.

3.2. Performance Guarantees of ULwE

The following theorem provides a regret upper bound for Algorithm 1:

Theorem 3.1. *In any case of nonstationary arrivals, the algorithm guarantees*

$$OPT \leq \left(1 + \frac{(1 + \min_{i \in [n]} c_i)\left(1 - e^{-1/\min_{i \in [n]} c_i}\right)}{1 - 1/e} \right) \mathbb{E}[ALG] + \tilde{O}(\sqrt{n|\Theta|T}).$$

When the arrivals are stationary, the algorithm guarantees

$$OPT \leq \mathbb{E}[ALG] + \tilde{O}(\sqrt{n|\Theta|T}).$$

The complete proof of Theorem 3.1 is proved in the Appendix, and we provide a sketch proof in Section 3.3. The Therorem establishes an expected regret bound of $\tilde{O}(\sqrt{n|\Theta|T})$ for near-i.i.d. arrival scenarios and provides a unified regret bound for nonstationary arrivals, combining sublinear regret with a constant CR. In the context of resource allocation, these results are significant.The sublinear regret of our algorithm indicates that the gap between its performance and the optimal strategy narrows over time. This demonstrates the algorithm's ability to adapt and improve, becoming more effective with longer usage. The constant CR highlights that our algorithm's regret is always within a fixed factor of the optimal strategy's regret, regardless of the problem's scale. This underlines the algorithm's consistent effectiveness and robustness, even with changing customer preferences. A remark on the behavior of the CR under the scenario when $\min_{i\in[n]} c_i$ approaches infinity, are provided in the Remark D.8.

3.3. Proof Sketch of Theorem 3.1

We provide an overview on the proof of Theorem 3.2, which is fully proved in Appendix D. First, we establish that ALG_{LP} always incurs a sublinear regret and a linear rate of resource consumption with high probability as long as no switch occurs (Proposition 3.2). Specifically, if ALG_{LP} runs uninterrupted for t iterations, the expected regret up to time t is bounded.

Theorem 3.2. *Suppose ALG_{LP} runs for t iterations without being switched, then the expected regret up to time t is upper bounded by*

$$16\sqrt{2|\Theta|nt\log(4|\Theta|t/\beta(n,T))}+\sqrt{5}n\log(2t/\beta(n,T))$$
$$+(2/t+2+\log t)\beta(n,T)(LP(\theta^*)+nt).$$

Moreover, each resource $i\in[n]$ has at least

$$\left[\frac{T-t}{T}c_i-\sqrt{8nt\log(2t/\beta(n,T))}\right]^+$$

remaining with probability at least $1-\beta(n,T)$.

We then demonstrate that under i.i.d. arrivals, the switch from ALG_{LP} to ALG_{ADV} does not occur with high probability (see Proposition D.13 and Proposition D.14). Consequently, we can apply Theorem 3.2, which assures sublinear regret without switch. This leads us to obtain $\tilde{O}(|\Theta|nT)$ regret under i.i.d. arrivals.

We further analyze our algorithm to address the case where the switch occurs, transitioning from ALG_{LP} to ALG_{ADV} at time t. Theorem 3.3 quantify the performance of the optimal algorithm from time $t+1$ to T , given the consumption of all resources is zero.

Theorem 3.3. *It holds that*

$$OPT\le\frac{(1+\min_{i\in[n]}c_i)\left(1-e^{-1/\min_{i\in[n]}c_i}\right)}{1-1/e}\mathbb{E}[ALG]$$
$$+\max_{i\in[n]}r_i(\sqrt{|\Theta|n}+1)\sqrt{2T\log(2T/\beta(n,T))}$$
$$+\max_{i\in[n]}r_i(1/T+\log T+1)/n.$$

This theorem is crucial as it sets an upper bound on the regret incurred by ALG_{ADV} under adversarial arrival conditions. By examining the collective performance of both ALG_{LP} and ALG_{ADV}, we derive a comprehensive regret bound in Theorem 3.1. Our analysis extends further to encompass nonstationary arrivals, showing that our algorithm not only achieves sublinear regret but also maintains a constant CR.

4. Numerical Studies

In this section, we conduct numerical experiments to assess the performance of ALG_{LP}, ALG_{ADV}, and the ULwE algorithm in a controlled setting where resources are continuously sold. The main goal is to evaluate these algorithms under dynamic conditions with limited resource availability, using $J^D(c,t)$ (Equation (1)). Based on insights from Golrezaei et al. (2014), we adjust our estimates of λ_l every $h=50$ periods to reflect shifts in customer arrivals while considering inventory levels and empirical data. Initially, based on the true rate μ, we generate data for 500 customers to estimate the initial λ_l values, which are used in the first h period cycles. Subsequently, every h interval, we update $\hat{\lambda}_l$s based on new data collected, using the formula $\hat{\lambda}_l=\frac{T}{t}\cdot\sum_{\tau=1}^{t}I_{\{customer\ at\ \tau\ is\ type\ l\}}$, where T is the total period, t is the current time, τ represents past time points, and $I_{\{customer\ at\ \tau\ is\ type\ l\}}$ is an indicator function that equals 1 if the customer of type l arrive at time τ, and 0 otherwise.

The experiments involve two customer types ($L=2$) and two resources ($n=2$), with customer purchase probabilities modeled through a logistic function. The ULwE algorithm starts with historical data (500 observations) to estimate initial Θ values, comparing against an optimal offline algorithm's maximum profit derived from known θ^* values. In this setup, Customer Type A exhibits a high purchase probability of 0.9, while Customer Type B maintains a consistent probability of 0.5, regardless of the resource. The total capacity of both resources matches the time horizon T, with revenues set at 1 and 1.5 units for Resource 1 and 2, respectively. Our analysis primarily assesses the ULwE algorithm's performance, particularly its regret in both i.i.d. and adversarial customer arrival scenarios, highlighting its adaptability across various operational contexts. Experiment's regret values represent the average outcomes of 100 independent experiments, ensuring robustness and reliability of the results.

4.1. Results under Near-IID Arrivals

In this subsection, we analyze experimental results under i.i.d. arrival conditions, where customer arrivals are stable and predictable. We simulate Customer Types A and B arriving at rates of $0.6T$ and $0.4T$, respectively, representing a near-i.i.d. environment with consistent arrival probabilities.

Figure 2 illustrates the regret trajectories over a total time period T of 500. The ADV algorithm (yellow line) initially maintains a stable regret, suggesting an optimal initial resource allocation. However, as time advances, an increase in regret is observed, a characteristic trait of the algorithm's greedy nature. This increase indicates a shift from optimal decisions to suboptimal ones due to the depletion of resources and consequent reduction in modified revenue. In

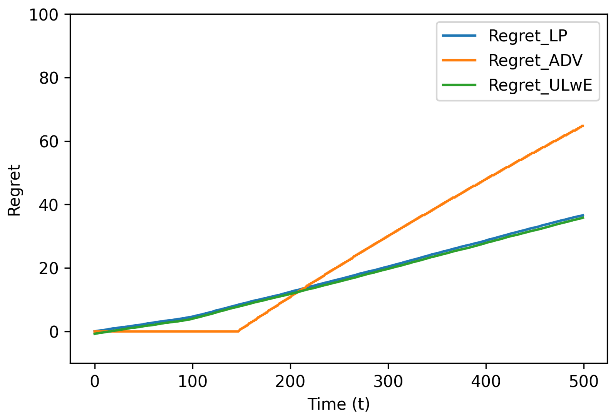

Figure 2. Regret over Time under i.i.d. Arrival

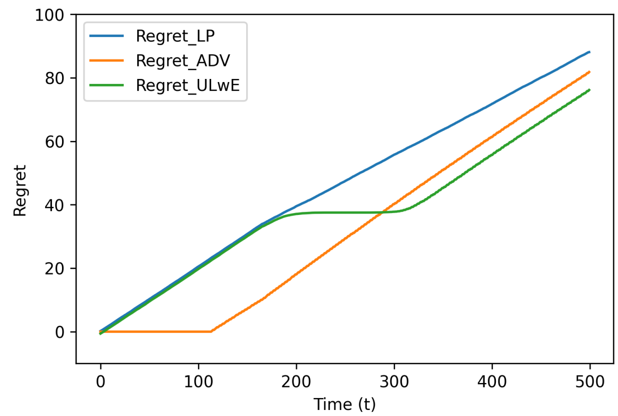

Figure 3. Regret over Time under Adversarial Arrival (ADV1)

contrast, the ALG_{LP} and ULwE algorithms (blue and green lines, respectively) exhibit a more gradual and consistent increase in regret. The near-overlapping of these lines indicates a similarity in their performance, with both algorithms showing a steady increase in revenue regret over time.

At $T = 500$, ALG_{ADV} has the highest cumulative regret, suggesting that the heuristic greedy algorithm encounters difficulties in adapting its resource allocation strategies effectively under i.i.d. arrivals. In contrast, both ALG_{LP} and the ULwE algorithm exhibit similar and improved performance compared to ALG_{ADV}, showing effective adaptation to i.i.d. conditions with lower cumulative regret. More details are provided in Appendix F.

4.2. Results under Adversarial Arrivals

In this subsection, we analyze how the algorithms perform in scenarios with nonstationary arrival patterns, wherein the probabilities of customer arrivals change over time. Specifically, we simulate a scenario in which Customer Types A and B exhibit varying arrival rates throughout different phases of the time period T to reflect dynamic shifts in customer behaviors. In our analysis of nonstationary environments, as shown in Figure 3, the regret trajectories of ALG_{LP} (blue line) and ALG_{ADV} (yellow line) were consistent with their performance under i.i.d. conditions. ALG_{LP} exhibited a steady increase in regret over time, while ALG_{ADV} showed initial stability followed by a sharp rise in the later stages. Interestingly, under nonstationary conditions, the regret of ALG_{ADV} did not surpass that of ALG_{LP}, indicating better performance in dynamic settings.

The ULwE algorithm (green line) displayed a distinct pattern. It initially follows the trend of ALG_{LP}, but after a switch point at around $1/3T$, it adopts a pattern similar to ALG_{ADV}—initially stable, then sharply increasing, indicating its ability to integrate the strengths of both the LP and ADV approaches. More details are provided in Appendix F.

4.3. Results under General Arrivals

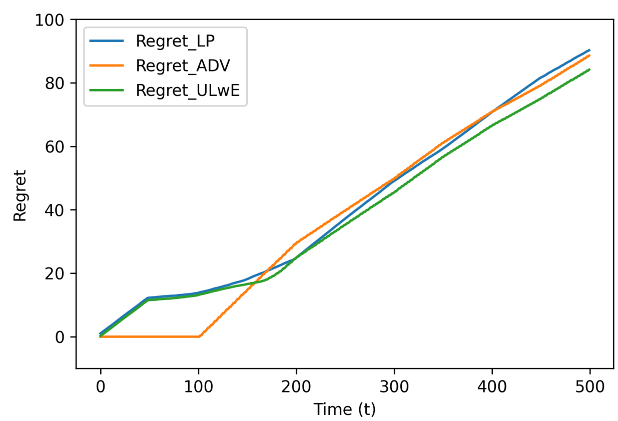

Figure 4. Regret over Time under Adversarial Arrival (ADV2)

This subsection evaluates the performance of each algorithms under various customer arrival scenarios (illustrated in Table 3), which vary in nonstationarity levels. The primary focus is on the ADV2 setting, indicative of high nonstationarity. Key findings highlight the robustness of the ULwE algorithm, especially in low nonstationarity settings like ADV1, attributed to its adaptive switching between ALG_{LP} and ALG_{ADV}. Although its performance dips in more nonstationary situations (e.g., ADV2), it still outperforms in stationary contexts.

In summary, ALG_{ADV} consistently sustains performance in nonstationary environments, while ALG_{LP}, limited by its static resource allocation, encounters challenges with fluctuating arrival rates and increased nonstationarity. ULwE's flexibility in resource allocation and its ability to adapt to changing conditions effectively minimize regret, highlighting its effectiveness in addressing nonstationary uncertainties.

5. Conclusion

In conclusion, we proposed ULwE algorithm based on CBwK to address the resource allocation problem under nonstationary environments. Our algorithm leverages contextual information to make informed decisions and adaptively balances exploration and exploitation to accommodate rapidly changing customer preferences. By assuming no prior knowledge of the per-period arrival process and considering the variability in arrival probabilities, our algorithm overcomes the constraints of stationary environments, thereby achieving efficient resource allocation.

While bandit algorithms have been used for resource allocation problems, we extend the existing literature by explicitly considering the variability in customer purchase probabilities. Compared to related works, our algorithm achieves a regret bound of approximately $\tilde{O}(\sqrt{n|\Theta|T})$ in the case of near-i.i.d. arrivals, providing sublinear regret and a constant CR under nonstationary arrivals. In addition to the theoretical analysis, our experiments compared ULwE algorithm with ALG_{LP} and ALG_{ADV}. The results consistently demonstrated that the ULwE algorithm outperforms these algorithms, achieving lower regret under nonstationary arrival patterns. Future research could explore the algorithm's applicability in other domains and enhance its performance in dynamic environments. In future work, we aim to develop a more computationally efficient version and even LP-free, which would significantly streamline the algorithm's practical application.

Acknowledgements

We would like to express their gratitude to the reviewing team for their constructive suggestions, which have significantly improved the paper's positioning and clarity of exposition.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agrawal, S. and Devanur, N. R. Linear Contextual Bandits with Knapsacks. *arXiv e-prints*, art. arXiv:1507.06738, July 2015. doi: 10.48550/arXiv.1507.06738.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. *arXiv e-prints*, art. arXiv:1506.03374, June 2015. doi: 10.48550/arXiv.1506.03374.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134, 2014.

Castiglioni, M., Celli, A., and Kroer, C. Online learning with knapsacks: the best of both worlds. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 17–23 Jul 2022.

Cetintas, S., Chen, D., Si, L., Shen, B., and Datbayev, Z. Forecasting counts of user visits for online display advertising with probabilistic latent class models. pp. 1217–1218, 07 2011. doi: 10.1145/2009916.2010127.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013.

Cheung, W. C., Ma, W., Simchi-Levi, D., and Wang, X. Inventory balancing with online learning. *Management Science*, 68(3):1776–1807, 2022.

Fahrbach, M., Huang, Z., Tao, R., and Zadimoghaddam, M. Edge-weighted online bipartite matching. *Journal of the ACM*, 69(6):1–35, 2022.

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

Fikioris, G. and Tardos, É. Approximately stationary bandits with knapsacks. *arXiv preprint arXiv:2302.14686*, 2023.

Golrezaei, N., Nazerzadeh, H., and Rusmevichientong, P. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014. doi: 10.1287/mnsc.2014.1939.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, 2019.

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. *J. ACM*, 69(6), nov 2022.

Jin, T., Liu, J., and Luo, H. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. *arXiv preprint arXiv:2302.13534*, 2023.

Karp, R. M., Vazirani, U. V., and Vazirani, V. V. An optimal algorithm for on-line bipartite matching. In *Proceedings of the Twenty-Second Annual ACM Symposium on Theory of Computing*, pp. 352–358, 1990.

Kesselheim, T. and Singla, S. Online learning with vector costs and bandits with knapsacks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*. PMLR, 09–12 Jul 2020.

Lai, H.-C., Shih, W.-Y., Huang, J.-L., and Chen, Y.-C. Predicting traffic of online advertising in real-time bidding systems from perspective of demand-side platforms. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3491–3498, 2016. doi: 10.1109/BigData.2016.7841012.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, S., Jiang, J., and Li, X. Non-stationary bandits with knapsacks. *ArXiv*, abs/2205.12427, 2022.

Luo, H., Tong, H., Zhang, M., and Zhang, Y. Improved high-probability regret for adversarial bandits with time-varying feedback graphs, 2023.

Lyu, L. and Cheung, W. C. Non-stationary bandits with knapsack problems with advice. *ArXiv*, abs/2302.04182, 2023.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized on-line matching. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pp. 264–273, 2005. doi: 10.1109/SFCS.2005.12.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

Richardson, M., Dominowska, E., and Ragno, R. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pp. 521–530, 2007.

Sivakumar, V., Zuo, S., and Banerjee, A. Smoothed adversarial linear contextual bandits with knapsacks. In *International Conference on Machine Learning*, 2022.

Slivkins, A. and Foster, D. J. Efficient contextual bandits with knapsacks via regression. *ArXiv*, abs/2211.07484, 2022.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Zhalechian, M., Keyvanshokooh, E., Shi, C., and Van Oyen, M. P. Online resource allocation with personalized learning. *Operations Research*, 70(4):2138–2161, 2022.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 467–475. PMLR, 2019.

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

A. Appendix: LP formulations

To capture the maximum achievable total reward, we define a static LP problem, denoted as LP(θ^*):

$$\begin{aligned} \text{LP}(\theta^*) = \max_{s_{ij}, i\in[n], j\in[L]} & \sum_{i\in[n]} r_i \sum_{j\in[L]} \lambda_j s_{ij} f_i(x^{(j)}, \theta_i^*) \\ \text{s.t.} & \sum_{j\in[L]} \lambda_j s_{ij} f_i(x^{(j)}, \theta_i^*) \le c_i,\ \forall i \in [n] \\ & \sum_{i\in[n]} s_{ij} = 1,\ \forall j \in [L] \\ & s_{ij} \ge 0,\ \forall i \in [n], j \in [L]. \end{aligned} \tag{7}$$

The coefficients in the LP are determined by the latent variables θ^*. The objective of this LP is to maximize the total reward obtained by assigning resources to customers based on their features and latent variables. The LP is subject to constraints that ensure consistency between budget limitations and resource assignment. The optimal solution to this LP is denoted as s^*.

We denote $\bar{U}$ the maximum optimal objective value of U^t, i.e.,

$$\begin{aligned} \bar{U} = \max_{s_{ij}, \theta_i\in\Theta_i, i\in[n], j\in[L]} & \sum_{i\in[n]} r_i \sum_{j\in[L]} \lambda_j s_{ij} \bar{f}_i(x^{(j)}, \theta_i) \\ \text{s.t.} & \sum_{j\in[L]} \lambda_j s_{ij} \bar{f}_i(x^{(j)}, \theta_i) \le c_i,\ \forall i \in [n] \\ & \sum_{i\in[n]} s_{ij} = 1,\ \forall j \in [L] \\ & s_{ij} \ge 0,\ \forall i \in [n], j \in [L]. \end{aligned} \tag{8}$$

B. Appendix: Proof of the Deterministic Upper Bound on the Optimal Revenue

In this subsection, our primary objective is to establish a theoretical upper bound on the optimal revenue that can be achieved. This involves a comprehensive analysis of the optimal revenue (denoted as $\overline{\text{OPT}}$), which serves as a benchmark against which the performance of various algorithms can be measured. By exploring the upper limits of achievable revenue, we aim to provide a clear framework for evaluating the efficiency and effectiveness of different resource allocation strategies. This analysis not only aids in quantifying the regret incurred by these algorithms but also offers insights into the potential for revenue maximization under varying operational constraints and market conditions.

Firstly, the resource allocation problem is formulated as follows: At time zero, the system has capacities c of n items and a finite time $t > 0$ to assign them.

Let $\sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*))$ denote the probability of item I is purchased by customers to time s. The term ${\mu_j}^t$ denotes the probability of encountering a customer of type j at time t, reflecting the non-stationary nature of customer arrivals. The function $f_I(x^{j^t}, \theta^*)$ captures the probability of a customer of type j with feature vector x^{j^t} purchasing resource I, given the latent variable θ^* associated with the resource. A consumption is realized at time s if $\sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*)) = 1$, in which case the system sells one item and receives revenue of r_I.

We introduce an indicator variable $\delta_{I,s}$, which equals 1 if resource I is assigned at time s, and 0 otherwise. The class of all non-anticipating policies, denoted by Π, must satisfy

$$\sum_{s=1}^{t} \sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*) \delta_{I,s} \le c_I$$

In addressing the challenge of optimizing resource allocation under constraints of customer behavior and time, we formulate a function that encapsulates the expected revenue over a continuous time horizon. The decision-making policy π guides the allocation of resources, determining each resource I being dynamically chosen based on customer features and the characteristics of the resources, as indicated by $\delta_{I,s}$.

Given a allocation policy $\pi \in \Pi$, an initial capacity $c > 0$, and a sales horizon $t > 0$, we denote the expected revenue by

$$J_\pi(c,t) \doteq \mathrm{E}_\pi \left[\sum_{s=1}^{t} r_I \sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*) \delta_{I,s} \right]$$

Here, r_I represents the revenue earned from resource I upon a customer's purchase, as determined by policy π and the assignment indicated by $\delta_{I,s}$. The problem is to find a decision-making policy π^* (if one exists) that maximizes the total expected revenue generated over $[0, t]$, denoted OPT. Equivalently,

$$\text{OPT} \doteq \sup_{\pi \in \Pi} J_\pi(c,t)$$

This formulation allows for a dynamic and adaptive approach to resource allocation, where the decision policy π can be optimized based on the varying probabilities of customer types and their purchasing behaviors over time. The objective OPT thus represents the total expected revenue, taking into account the fluctuating customer landscape and the inherent uncertainties in customer preferences and behaviors.

In our pursuit to devise an optimal strategy for resource allocation in a customer interaction system, we propose a deterministic linear programming model, denoted as J^D. This model is designed to maximize the deterministic revenue while adhering to the constraints of resource capacities and customer arrival patterns. The objective function aims to maximize the total deterministic revenue overall resources, customer types, and time periods.

$$\begin{aligned} J^D(c,t) = \max_{s_{ij}, i \in [n], j \in [L]} & \sum_{i=1}^{n} \sum_{j=1}^{L} \sum_{s=1}^{t} r_i s_{ij}^s f_i(x^j, \theta^*) \\ \text{s.t.} & \sum_{j=1}^{L} \sum_{s=1}^{t} {\mu_j}^s s_{ij}^s f_i(x^j, \theta^*) \le c_i, \forall i \in [n] \\ & \sum_{i=1}^{n} \sum_{s=1}^{t} s_{ij}^s = \sum_{s=1}^{t} {\mu_j}^s, \forall j \in [L] \\ & s_{ij}^s \ge 0, \forall i \in [n], \forall j \in [L] \end{aligned}$$

All the variables and parameters in this LP are deterministic. They are either constants known a priori or decision variables whose values are to be determined by solving the LP without any randomness involved. The first constraint ensures that the expected number of resource allocations does not exceed the capacity c_i of each resource i. This constraint is crucial for maintaining a balance between maximizing revenue and not overcommitting the available resources. The second constraint guarantees that the total allocations for each customer type j over the time horizon equal the expected number of arrivals λ_j of that customer type. This aligns the resource allocation with the anticipated customer demand.

Finally, the non-negativity constraint ensures that the decision variables remain feasible, reflecting the reality that negative resource allocation is not possible.

Proposition B.1. *For all* $0 \le c < +\infty$ *and* $0 \le t < +\infty$*,*

$$OPT \le J^D(c,t)$$

Proof. Initially, the problem is formulated as finding the best policy π to maximize the expected sum of returns, subject to a cost constraint:

$$\begin{aligned} \text{OPT} \doteq \max_{\pi} \quad & \mathrm{E}_\pi \left[\sum_{s=1}^{t} r_I \sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*) \delta_{I,s} \right] \\ \text{s.t.} \quad & \sum_{s=1}^{t} \sum_{j=1}^{L} {\mu_j}^s f_I(x^{j^s}, \theta^*)) \le c_I \end{aligned}$$

Here we introduce binary decision variables π_{ij} to represent the allocation of resources to each customer at each time point $s > 0$. We add a constraint $\sum_{i=1}^{n} \pi_{ij}^s = 1$, which represents each customer can only be assigned one item at each time point. The problem thus becomes a binary integer programming problem:

$$
\begin{aligned}
\text{OPT} = \max_{\pi_{ij}} & \sum_{i=1}^{n} \sum_{j=1}^{L} \sum_{s=1}^{t} r_i {\mu_j}^s \pi_{ij}^s f_i(x^{j^s}, \theta^*) \\
\text{s.t.} & \sum_{s=1}^{t} \sum_{j=1}^{L} {\mu_j}^s \pi_{ij}^s f_i(x^{j^s}, \theta^*)) \leq c_i \quad \forall i \in [n] \\
& \sum_{i=1}^{n} \pi_{ij}^s = 1 \\
& \pi_{ij}^s = \{0, 1\}
\end{aligned}
$$

To establish the relationship between the original optimization problem OPT and $J^D(c, t)$, we begin to get a new LP problem OPT_1 by relaxing the binary constraint on π_{ij}^s, allowing it to take any value between 0 and 1. This relaxation naturally leads to $OPT_1 \geq$ OPT, as OPT_1 includes all feasible solutions of OPT and potentially more. We then introduce a new variable $m_{ij}^s = {\mu_j}^s \pi_{ij}^s$, noting that since ${\mu_j}^s$ is also between 0 and 1, it follows that $0 \leq m_{ij}^s \leq 1$.

The first constraint of OPT_1, $\sum_{s=0}^{t} \sum_{j=1}^{L} m_{ij}^s f_i(x^{j^s}, \theta^*)) \leq c_i$, is tighter than the first constraint of $J^D(c, t)$, $\sum_{s=0}^{t} \sum_{j=1}^{L} {\mu_j}^s s_{ij}^s f_i(x^{j^s}, \theta^*)) \leq c_i$, as $\sum_{s=0}^{t} \sum_{j=1}^{L} m_{ij}^s f_i(x^{j^s}, \theta^*))$ is larger than $\sum_{s=0}^{t} \sum_{j=1}^{L} {\mu_j}^s s_{ij}^s f_i(x^{j^s}, \theta^*))$.

The second constraint of OPT_1, $\sum_{i=1}^{n} m_{ij}^s = \sum_{i=1}^{n} {\mu_j}^s \pi_{ij}^s = {\mu_j}^s$, applies to each time period, making it more stringent than the overall time constraint in $J^D(c, t)$,$\sum_{i=1}^{n} \sum_{s=1}^{t} s_{ij}^s = \sum_{s=1}^{t} {\mu_j}^s$. Consequently, we conclude that OPT $\leq OPT_1 \leq J^D(c, t)$, completing the proof. □

Having rigorously established the inequality OPT $\leq J^D(c, t)$, we have effectively demonstrated that $J^D(c, t)$ serves as an upper bound on the optimal revenue, which we denote as $\overline{\text{OPT}}$. This upper bound is crucial as it provides a benchmark against which the performance of various algorithms can be measured. In light of this, we are now positioned to define the concept of 'regret' in our context. Specifically, regret can be quantitatively expressed as the difference between the upper bound of the optimal revenue and the actual optimal revenue achieved, formulated as

$$
\text{Regret} = \overline{\text{OPT}} - \text{ALG}
$$

This definition of regret is instrumental in evaluating the efficiency of different algorithms. By measuring how closely an algorithm's performance approaches the theoretical upper limit of revenue, we can assess its effectiveness in resource allocation and decision-making under various operational scenarios.

C. Appendix: Technical Lemmas

Lemma C.1. *(Azuma-Hoeffding inequality) Let $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots \subset \mathcal{G}_n$ be a filtration* **G***, and $X_0, \ldots, X_n$ a martingale associated with* **G***, and $|X_i - X_{i-1}| \leq c_i$, $\forall\, i = 1, \ldots, n$ almost surely. Then, it holds that*

$$
\mathbb{P}\left[|X_n - X_0| > \epsilon\right] \leq 2 \exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{n} c_k^2}\right),
$$

for all $\epsilon > 0$.

Lemma C.2. *((Badanidiyuru et al., 2014)) Let $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots \subset \mathcal{G}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{G}_i$-measurable, $\mathbb{E}[X_i|\mathcal{G}_{i-1}] = 0$ and $|X_i| \leq c$ for all $i \in [n]$ and some $c > 0$. Then with probability at least $1 - \delta$ it holds that*

$$
\left|\sum_{i=1}^{n} X_i\right| \leq \sqrt{4 \sum_{i=1}^{n} \mathbb{E}[X_i^2|\mathcal{G}_{i-1}] \log(2n/\delta) + 5c^2 \log^2(2n/\delta)}.
$$

Proof. First notice for all $i \in [n], j \in [L], t \in [T]$, we have

$$\begin{aligned}\mathbb{E}\left[\mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) | \mathcal{F}_{t-1}\right] =& \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(i,j,t)|\mathcal{F}_{t-1}\right] \bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right] \\ =& \mathbb{E}\left[\mathbb{I}(i,j,t)|\mathcal{F}_{t-1}\right] \mathbb{E}\left[\bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right] \\ =& \mathbb{E}\left[\mathbb{I}(i,j,t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right],\end{aligned}$$

and

$$\left|\sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})\right| \leq 1.$$

Let $X_0 = 0$, $X_t - X_{t-1} = \sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})$, then we know $\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = 0$ for all $t \in [T]$. To use Lemma C.2, we need to estimate $V_T := \sum_{t=1}^{T} \mathbb{E}\left[X_t^2|\mathcal{F}_{t-1}\right]$. To be specific, we have

$$\begin{aligned}V_T =& \sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})\right)^2 \middle| \mathcal{F}_{t-1}\right] \\ =& \sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)^2\bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1}\right] \\ &+ \sum_{t=1}^{T}\underbrace{\mathbb{E}\left[\sum_{j\neq j'\in[L]}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)(\mathbb{I}(i,j',t) - \mu_j \bar{s}_{ij'}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})\bar{f}_i(x^{(j')}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right]}_{\leq 0} \\ \leq& \sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)^2\bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1}\right] \\ =& \sum_{t=1}^{T}\left(\mathbb{E}\left[\sum_{j=1}^{L}\mathbb{I}(i,j,t)^2\bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1}\right] - 2\mathbb{E}\left[\sum_{j=1}^{L}\mathbb{I}(i,j,t)\mu_j \bar{s}_{ij}^t\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right]\right) \\ &+ \sum_{t=1}^{T}\mathbb{E}\left[\mu_j^2(\bar{s}_{ij}^t)^2\bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1}\right] \\ \leq& \sum_{t=1}^{T}\left(\mathbb{E}\left[\sum_{j=1}^{L}\mathbb{I}(i,j,t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right] - 2\mathbb{E}\left[\sum_{j=1}^{L}\mathbb{I}(i,j,t)\mu_j \bar{s}_{ij}^t\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right]\right) \\ &+ \sum_{t=1}^{T}\mathbb{E}\left[\mu_j^2(\bar{s}_{ij}^t)^2\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right] \\ =& \sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{L}\mu_j \bar{s}_{ij}^t(1 - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right] \\ \leq& \sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{L}\mu_j \bar{s}_{ij}^t\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1}\right].\end{aligned}$$

The first inequality is due to the following calculation: given any $j \neq j' \in [L]$,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbb{I}(i,j,t)-\mu_j\bar{s}_{ij}^t)(\mathbb{I}(i,j',t)-\mu_j\bar{s}_{ij'}^t)\bar{f}_i(x^{(j)},\Omega_i^{t-1})\bar{f}_i(x^{(j')},\Omega_i^{t-1})\middle|\mathcal{F}_{t-1}\right]\\
=&(\mu_j\bar{s}_{ij}^t(1-\mu_j\bar{s}_{ij}^t)(-\mu_{j'}\bar{s}_{ij'}^t)+\mu_{j'}\bar{s}_{ij'}^t(-\mu_j\bar{s}_{ij}^t)(1-\mu_{j'}\bar{s}_{ij'}^t)\\
+&(\mu_j(1-\bar{s}_{ij}^t)+\mu_{j'}(1-\bar{s}_{ij'}^t)+(1-\mu_j-\mu_{j'}))\mu_j\bar{s}_{ij}^t\mu_{j'}\bar{s}_{ij'}^t)\cdot\bar{f}_i(x^{(j)},\Omega_i^{t-1})\bar{f}_i(x^{(j')},\Omega_i^{t-1})\\
=&-\mu_j\bar{s}_{ij}^t\mu_{j'}\bar{s}_{ij'}^t\bar{f}_i(x^{(j)},\Omega_i^{t-1})\bar{f}_i(x^{(j')},\Omega_i^{t-1})\leq 0.
\end{aligned}$$

So, by using Lemma C.2, for any $0<\delta<1$, we obtain with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{t=1}^T\sum_{j=1}^L\left(\mathbb{I}(i,j,t)-\mu_j\bar{s}_{ij}^t\right)\bar{f}_i(x^{(j)},\Omega_i^{t-1})\right|\\
\leq&\sqrt{4\sum_{t=1}^T\mathbb{E}\left[\sum_{j=1}^L\mu_j\bar{s}_{ij}^t\bar{f}_i(x^{(j)},\Omega_i^{t-1})\middle|\mathcal{F}_{t-1}\right]\log(2T/\delta)+5\log^2(2T/\delta)}\\
\leq&\sqrt{4\sum_{t=1}^T\mathbb{E}\left[\sum_{j=1}^L\mu_j\bar{s}_{ij}^t\bar{f}_i(x^{(j)},\Omega_i^{t-1})\middle|\mathcal{F}_{t-1}\right]\log(2T/\delta)}+\sqrt{5}\log(2T/\delta),
\end{aligned}$$

where in the second inequality we use the fact $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for any $a,b,>0$. Therefore, we conclude with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{i=1}^n\left|\sum_{t=1}^T\sum_{j=1}^L\left(\mathbb{I}(i,j,t)-\mu_j\bar{s}_{ij}^t\right)\bar{f}_i(x^{(j)},\Omega_i^{t-1})\right|\\
\leq&\sum_{i=1}^n\sqrt{4\sum_{t=1}^T\mathbb{E}\left[\sum_{j=1}^L\mu_j\bar{s}_{ij}^t\bar{f}_i(x^{(j)},\Omega_i^{t-1})\middle|\mathcal{F}_{t-1}\right]\log(2T/\delta)}+\sqrt{5}n\log(2T/\delta)\\
\leq&\sqrt{4n\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\left[\sum_{j=1}^L\mu_j\bar{s}_{ij}^t\bar{f}_i(x^{(j)},\Omega_i^{t-1})\middle|\mathcal{F}_{t-1}\right]\log(2T/\delta)}+\sqrt{5}n\log(2T/\delta)\\
\leq&\sqrt{4nT\log(2T/\delta)}+\sqrt{5}n\log(2T/\delta),
\end{aligned}$$

where in the second inequality we use the Cauchy-Schwarz inequality. □

Corollary C.3. *After the modification in the removal process of θ,*

$$\mathbb{P}[\mathcal{E}]\geq 1-(1+\log T)\beta(n,T).$$

Corollary C.3 is directly turned out by Proposition D.3.

Lemma C.4. *For any $0<\delta<1$ and $t\in[T]$, it holds that*

$$\mathbb{P}\left[\forall\,\theta\in\Theta:\left|\sum_{l=1}^t\sum_{i=1}^n r_i(s_{iJ^l}(\theta)-\bar{s}_{iJ^l}^l-\sum_{j=1}^L\mu_j^l(s_{ij}(\theta)-\bar{s}_{ij}^l))f_{ij}(x^{(j)},\theta)\right|>\max_{i\in[n]}r_i\sqrt{8t\log(2|\Theta|/\delta)}\right]\leq\delta.$$

Proof. First note $\mathbb{E}[(s_{iJ^t}(\theta)-\bar{s}_{iJ^t}^t)f_{iJ^t}(x^{(J^t)},\theta)|\mathcal{F}_{t-1}]=\sum_{j=1}^L\mu_j^t(s_{ij}(\theta)-\bar{s}_{ij}^t)f_i(x^{(J^t)},\theta)$ and $\left|\sum_{i=1}^n r_i\left(s_{iJ^l}(\theta)-\bar{s}_{iJ^t}^t-\sum_{j=1}^L\mu_j^t(s_{ij}(\theta)-\bar{s}_{ij}^t)\right)f_i(x^{(j)},\theta)\right|\leq 2\max_{i\in[n]}r_i$ for all $t\in[T]$, so we define $X_0=0,X_l-$

$X_{l-1} = \sum_{i=1}^{n} r_i \left(s_{iJ^l}(\theta) - \bar{s}^l_{iJ^l} - \sum_{j=1}^{L} \mu^l_j (s_{ij}(\theta) - \bar{s}^l_{ij}) \right) f_i(x^{(j)}, \theta)$ so that $X_0, \ldots, X_t$ is a martingale associated with the filtration $\mathcal{F}_0, \ldots, \mathcal{F}_{t-1}$. So by using the Azuma-Hoeffding inequality (c.f. Lemma C.1) we have for any $\theta \in \Theta$,

$$\mathbb{P}\left[\left|\sum_{l=1}^{t}\sum_{i=1}^{n} r_i \left(s_{iJ^l}(\theta) - \bar{s}^l_{iJ^l} - \sum_{j=1}^{L} \mu^l_j (s_{ij}(\theta) - \bar{s}^l_{ij}) \right) f_i(x^{(j)}, \theta)\right| > \epsilon\right] \le 2e^{-\epsilon^2/(8t(\max_{i\in[n]} r_i)^2)}.$$

By letting $\delta/|\Theta| = 2e^{-\epsilon^2/(8t(\max_{i\in[n]} r_i)^2)}$ and using the union bound we obtain the desired result. □

Lemma C.5. *For any $0 < \delta < 1$ and $t \in [T]$, it holds that*

$$\mathbb{P}\left[\left|\sum_{l=1}^{t}\left(\bar{s}^l_{iJ^l} f_i(x^{(J^l)}, \Omega^{l-1}) - \sum_{j=1}^{L} \mu_j \bar{s}^l_{ij} f_i(x^{(j)}, \Omega^{l-1})\right)\right| > \sqrt{2t\log(2/\delta)}\right] \le \delta.$$

D. Appendix: Omitted Proofs from Section 3

D.1. Proofs under Stationary Arrivals

We point out the expected regret in period t is (for now ignore events at the end of the horizon)

$$\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s^*_{ij} - \bar{s}^t_{ij}) f_i(x^{(j)}, \theta^*_i).$$

The following proposition states that the regret can be bounded from above by the size of the confidence intervals of purchase probabilities in period t.

Proposition D.1. *Suppose in each period t, if $\bar{f}_i(x^{(j)}, \Omega^{t-1}_i) \ge f_i(x^{(j)}, \theta^*_i)$ for all $i \in [n], j \in [L]$, then*

$$\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s^*_{ij} - \bar{s}^t_{ij}) f_i(x^{(j)}, \theta^*_i) \le \sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j \bar{s}^t_{ij} (\bar{f}_i(x^{(j)}, \Omega^{t-1}_i) - f_i(x^{(j)}, \theta^*_i)).$$

Proof.

$$\begin{aligned}
&\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s^*_{ij} - \bar{s}^t_{ij}) f_i(x^{(j)}, \theta^*_i) \\
=&\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j s^*_{ij} f_i(x^{(j)}, \theta^*_i) - \sum_{i=1}^{n} r_i \left[\sum_{j=1}^{L} \mu_j \bar{s}^t_{ij} \bar{f}_i(x^{(j)}, \Omega^{t-1}_i) - \sum_{j=1}^{L} \mu_j \bar{s}^t_{ij} (\bar{f}_i(x^{(j)}, \Omega^{t-1}_i) - f_i(x^{(j)}, \theta^*_i))\right] \\
=&LP(\theta^*) - U^t + \sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j \bar{s}^t_{ij} (\bar{f}_i(x^{(j)}, \Omega^{t-1}_i) - f_i(x^{(j)}, \theta^*_i)) \\
\le&\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j \bar{s}^t_{ij} (\bar{f}_i(x^{(j)}, \Omega^{t-1}_i) - f_i(x^{(j)}, \theta^*_i)).
\end{aligned}$$

The last inequality follows from the condition that the values of "purchase probabilities" in U^t (2) are at least those in LP(θ^*) (7). To see this, we define $\bar{s}_{ij} = s^*_{ij} \cdot \eta_{ij}$ where $\eta_{ij} := f_i(x^{(j)}, \theta^*_i)/f_i(x^{(j)}, \Omega^{t-1}_i)$ if $f_i(x^{(j)}, \theta^*_i) \neq 0$. Then, since $\eta_{ij} \le 1$ for all i, j, $\bar{s}_{ij}$ is always a feasible solution for U^t with revenue $\sum_{i\in[n]} r_i \sum_{j\in[L]} \lambda_j s^*_{ij} f_i(x^{(j)}, \theta^*_i)$. This is because by definition we have $\sum_{j\in[L]} \lambda_j s^*_{ij} f_i(x^{(j)}, \theta^*_i) \le c_i$ for all i and suppose some null resource exists (i.e., representing the "no-click" event) then we could allocate the additional allocation weights to a null resource with zero revenue). This proves $U^t \ge$ LP(θ^*). □

Proposition D.1 provides an alternative perspective on the regret incurred in period t. Specifically, the algorithm selects "arm" i under context $x^{(j)}$ with probability $\mu_j \bar{s}_{ij}^t$. In this case, the algorithm experiences regret equal to $\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) - f_i(x^{(j)}, \theta_i^*)$.

We next provide a direct result from the Azuma-Hoeffding inequality (see Lemma C.1).

Lemma D.2. *For any* $0 < \delta < 1$, $t \in [T]$, *suppose all* I^t, J^t *are given,*

$$\mathbb{P}\left[\left|\sum_{l=1}^{t}(f_{I^l}(x^{(J^l)}, \theta_{I^l}^*) - a^l)\right| > \sqrt{2t\log(2/\delta)}\right] \leq \delta.$$

Proof. First note $\mathbb{E}\left[a^l | \mathcal{F}_{l-1}\right] = f_{I^l}(x^l, \theta_{I^l}^*)$ and $|f_{I^l}(x^{(J^l)}, \theta_{I^l}^*) - a^l| \leq 1$ for all $l \in [t]$. Let $X_0 = 0, X_l - X_{l-1} = f_{I^l}(x^l, \theta_{I^l}^*) - a^l$ for all $l \in [t]$ we know $X_0, \ldots, X_l$ is a martingale associated with the filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_{l-1}$. Thus using the Azuma-Hoeffding inequality (c.f. Lemma C.1) we have for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{l=1}^{t} r_{I^t}(f_{I^l}(x^l, \theta_{I_l}^*) - a^l)\right| > \epsilon\right] \leq 2e^{-\epsilon^2/(2t)}.$$

Then the desired result is obtained by setting ϵ such that $\delta = 2e^{-\epsilon^2/(2t)}$. □

Proposition D.3 shows the probabilistic event $\mathcal{E} := \forall\, i \in [n], j \in [L], t \in [T] : \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \geq f_i(x^{(j)}, \theta_i^*)$ indeed happens with high probability. The event $\mathcal{E}$ ensures that the maximum purchase probability for resource i based on a set Ω of valid latent variables is greater than the true probability. Once the probabilistic event $\mathcal{E}$ has been established to occur with high probability, analyzing the revenue regret under different conditions becomes more convenient.

Proposition D.3. $\mathbb{P}\left[\mathcal{E}\right] \geq 1 - (1 + \log T)\beta(n, T)$.

Proof. Since

$$\mathbb{P}\left[\forall\, i \in [n], j \in [L], t \in [T] : \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \geq f_i(x^{(j)}, \theta_i^*)\right] \geq \mathbb{P}\left[\forall\, i \in [n], t \in [T] : \theta_i^* \in \Omega_i^t\right],$$

so it is sufficient to lower bound $\mathbb{P}\left[\forall\, i \in [n], t \in [T] : \theta_i^* \in \Omega_i^t\right]$.

Consider the probabilistic event

$$A(t) = \left\{\left|\sum_{l=1}^{t} f_{I^l}(x^{(J^l)}, \theta_{I^l}^*) - a^l\right| > \sqrt{2t\log(2t/\beta(n,T))}\right\}.$$

We know from Lemma D.2 that

$$\mathbb{P}\left[A(t)\right] \leq \beta(n, T)/t.$$

From step 3 of the algorithm and then using the union bound, we have

$$\begin{aligned}
\mathbb{P}\left[\forall\, i \in [n], t \in [T] : \theta_i^* \in \Omega_i^t\right] =& \mathbb{P}\left[\left|\sum_{t' \in D_{I^t}^t(\theta_i^*)} f_{I^{t'}}(x^{t'}, \theta_i^*) - a^{t'}\right| \leq \sqrt{2t\log(2t/\beta(n,T))}\right] \\
\geq& \mathbb{P}\left[\left|\sum_{l=1}^{t} f_{I^l}(x^{(J^l)}, \theta_{I^l}^*) - a^l\right| \leq \sqrt{2t\log(2t/\beta(n,T))}\right] \\
\geq& 1 - \mathbb{P}\left[\cup_{t=1}^{T} A(t)\right] \\
\geq& 1 - \sum_{t=1}^{T} \beta(n, T)/t \\
\geq& 1 - (\log T + 1)\beta(n, T),
\end{aligned}$$

where in the last inequality we use the fact $\sum_{t=1}^{T} \frac{1}{t} \leq \log T + 1$.

□

Proposition D.4 establishes an upper bound on the expected regret resulting from the deviation of revenue. It quantifies the relationship between the regret and the revenue shortfall caused by the algorithm's decisions. By considering this upper bound, we can gain insights into the potential regret incurred due to revenue variations.

Proposition D.4. *We have*

$$\mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s_{ij}^* - \bar{s}_{ij}^t) f_i(x^{(j)}, \theta_i^*)}_{(*)}\right]$$
$$\leq \max_{i\in[n]} r_i \left(\sqrt{n \max_{i\in[n]} |\Theta_i|} + 1\right) \sqrt{8T\log(2T/\beta(n,T))} + \bar{U}/T\beta(n,T) + LP(\theta^*)\beta(n,T)(\log T + 1).$$

Proof. From Proposition D.1, given $\mathcal{E}$, it is sufficient to upper bound the term

$$\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t (\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) - f_i(x^{(j)}, \theta_i^*))}_{(**)}$$
$$= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mathbb{I}(i,j,t)(\bar{f}_i(x^{(j)}, \Omega_i^{t-1}) - f_i(x^{(j)}, \theta_i^*)) \middle| \mathcal{F}_{t-1}\right]$$
$$= \sum_{t=1}^{T} \mathbb{E}\left[r_{I^t}(\bar{f}_{I^t}(x^{(J^t)}, \Omega_{I^t}^{t-1}) - f_{I^t}(x^{(J^t)}, \theta_{I^t}^*)) \middle| \mathcal{F}_{t-1}\right]$$
$$\leq \max_{i\in[n]} r_i \cdot \mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\left(\bar{f}_{I^t}(x^{(J^t)}, \Omega_{I^t}^{t-1}) - a^t\right)}_{(***)} + \underbrace{\sum_{t=1}^{T}\left(a^t - f_{I^t}(x^{(J^t)}, \theta_{I^t}^*)\right)}_{(****)} \middle| \mathcal{F}_{t-1}\right],$$

where the first inequality holds because of $\mathbb{E}[\mathbb{I}(i,j,t)|\mathcal{F}_t] = \mu_t \bar{s}_{ij}^t$ and law of iterated expectation. Next, from step 3 of the algorithm and the Cauchy-Schwarz inequality, it holds almost surely

$$|(***)| = \left|\sum_{i=1}^{n}\sum_{\theta_i\in\Theta_i}\sum_{t'\in D_i^T(\theta_i)} \left(f_i(x^{(J^{t'})}, \theta_i) - a^{t'}\right)\right|$$
$$\leq \sum_{i=1}^{n}\sum_{\theta_i\in\Theta_i} \sqrt{2|D_i^T(\theta_i)|\log(2T/\beta(n,T))}$$
$$\leq \sqrt{n\max_{i\in[n]}|\Theta_i| \sum_{i=1}^{n}\sum_{\theta_i\in\Theta_i} 2|D_i^T(\theta_i)|\log(2T/\beta(n,T))}$$
$$= \sqrt{2n\max_{i\in[n]}|\Theta_i| T\log(2T/\beta(n,T))}.$$

Moreover, from D.2 we know

$$\mathbb{P}\left[|(****)| \leq \sqrt{2T\log(2T/\beta(n,T))}\right] \geq 1 - \beta(n,T)/T.$$

Therefore, let $\alpha = \max_{i\in[n]} r_i \left(\sqrt{n \max_{i\in[n]} |\Theta_i|} + 1\right) \sqrt{8T\log(2T/\beta(n,T))}$, we obtain

$$\mathbb{P}[(**) \leq \alpha | \mathcal{E}] \geq 1 - \beta(n,T)/T.$$

Therefore, noticing $\mathbb{P}[(**) > \alpha, \mathcal{E}] \leq \mathbb{P}[(**) > \alpha|\mathcal{E}] \leq \beta(n,T)/T$,

$$\begin{aligned}
\mathbb{E}[(*)] =& \mathbb{E}[(*)|\mathcal{E}]\mathbb{P}[\mathcal{E}] + \mathbb{E}[(*)|\mathcal{E}^c]\mathbb{P}[\mathcal{E}^c] \\
\leq& \mathbb{E}[(**)|\mathcal{E}] + LP(\theta^*)\beta(n,T)(\log T + 1) \\
=& \mathbb{E}[(**)|(**) \leq \alpha, \mathcal{E}]\mathbb{P}[(**) \leq \alpha, \mathcal{E}] + \mathbb{E}[(**)|(**) > \alpha, \mathcal{E}]\mathbb{P}[(**) > \alpha, \mathcal{E}] \\
&+ LP(\theta^*)\beta(n,T)(\log T + 1) \\
\leq& \alpha + \bar{U}\beta(n,T)/T + LP(\theta^*)\beta(n,T)(\log T + 1).
\end{aligned}$$

□

Lemma D.5 provides a crucial concentration estimate in analyzing the regret incurred by the violation of the inventory constraints.

Lemma D.5. *For any* $0 < \delta < 1$*, it holds that,*

$$\mathbb{P}\left[\sum_{i=1}^{n}\left|\sum_{t=1}^{T}\sum_{j=1}^{L}\left(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t\right)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})\right| > \sqrt{4nT\log(2T/\delta)} + \sqrt{5}n\log(2T/\delta)\right] \leq \delta.$$

Proof. First notice for all $i \in [n], j \in [L], t \in [T]$, we have

$$\begin{aligned}
\mathbb{E}\left[\mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right] =& \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(i,j,t)|\mathcal{F}_{t-1}\right]\bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right] \\
=& \mathbb{E}\left[\mathbb{I}(i,j,t)|\mathcal{F}_{t-1}\right]\mathbb{E}\left[\bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right] \\
=& \mathbb{E}\left[\mathbb{I}(i,j,t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})|\mathcal{F}_{t-1}\right],
\end{aligned}$$

and

$$\left|\sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})\right| \leq 1.$$

Let $X_0 = 0$, $X_t - X_{t-1} = \sum_{j=1}^{L}(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})$, then we know $\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = 0$ for all $t \in [T]$. To use Lemma C.2, we need to estimate $V_T := \sum_{t=1}^{T}\mathbb{E}\left[X_t^2|\mathcal{F}_{t-1}\right]$. To be specific, we have

$$
\begin{aligned}
V_T =& \sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_{j=1}^{L} (\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \right)^2 \middle| \mathcal{F}_{t-1} \right] \\
=& \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} (\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)^2 \bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1} \right] \\
&+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\sum_{j \neq j' \in [L]} (\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)(\mathbb{I}(i,j',t) - \mu_j \bar{s}_{ij'}^t) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \bar{f}_i(x^{(j')}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right]}_{\leq 0} \\
\leq& \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} (\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)^2 \bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1} \right] \\
=& \sum_{t=1}^{T} \left(\mathbb{E}\left[\sum_{j=1}^{L} \mathbb{I}(i,j,t)^2 \bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1} \right] - 2\mathbb{E}\left[\sum_{j=1}^{L} \mathbb{I}(i,j,t) \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \right) \\
&+ \sum_{t=1}^{T} \mathbb{E}\left[\mu_j^2 (\bar{s}_{ij}^t)^2 \bar{f}_i(x^{(j)}, \Omega_i^{t-1})^2 \middle| \mathcal{F}_{t-1} \right] \\
\leq& \sum_{t=1}^{T} \left(\mathbb{E}\left[\sum_{j=1}^{L} \mathbb{I}(i,j,t) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] - 2\mathbb{E}\left[\sum_{j=1}^{L} \mathbb{I}(i,j,t) \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \right) \\
&+ \sum_{t=1}^{T} \mathbb{E}\left[\mu_j^2 (\bar{s}_{ij}^t)^2 \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \\
=& \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t (1 - \mu_j \bar{s}_{ij}^t) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \\
\leq& \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right].
\end{aligned}
$$

The first inequality is due to the following calculation: given any $j \neq j' \in [L]$,

$$
\begin{aligned}
&\mathbb{E}\left[(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)(\mathbb{I}(i,j',t) - \mu_j \bar{s}_{ij'}^t) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \bar{f}_i(x^{(j')}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \\
=& (\mu_j \bar{s}_{ij}^t (1 - \mu_j \bar{s}_{ij}^t)(-\mu_{j'} \bar{s}_{ij'}^t) + \mu_{j'} \bar{s}_{ij'}^t (-\mu_j \bar{s}_{ij}^t)(1 - \mu_{j'} \bar{s}_{ij'}^t) \\
&+ (\mu_j (1 - \bar{s}_{ij}^t) + \mu_{j'} (1 - \bar{s}_{ij'}^t) + (1 - \mu_j - \mu_{j'})) \mu_j \bar{s}_{ij}^t \mu_{j'} \bar{s}_{ij'}^t) \cdot \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \bar{f}_i(x^{(j')}, \Omega_i^{t-1}) \\
=& -\mu_j \bar{s}_{ij}^t \mu_{j'} \bar{s}_{ij'}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \bar{f}_i(x^{(j')}, \Omega_i^{t-1}) \leq 0.
\end{aligned}
$$

So, by using Lemma C.2, for any $0 < \delta < 1$, we obtain with probability at least $1 - \delta$,

$$
\begin{aligned}
&\left| \sum_{t=1}^{T} \sum_{j=1}^{L} \left(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t \right) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \right| \\
&\le \sqrt{4 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \log(2T/\delta) + 5 \log^2(2T/\delta)} \\
&\le \sqrt{4 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \log(2T/\delta)} + \sqrt{5} \log(2T/\delta),
\end{aligned}
$$

where in the second inequality we use the fact $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b, > 0$. Therefore, we conclude with probability at least $1 - \delta$,

$$
\begin{aligned}
&\sum_{i=1}^{n} \left| \sum_{t=1}^{T} \sum_{j=1}^{L} \left(\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t \right) \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \right| \\
&\le \sum_{i=1}^{n} \sqrt{4 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \log(2T/\delta)} + \sqrt{5} n \log(2T/\delta) \\
&\le \sqrt{4n \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \middle| \mathcal{F}_{t-1} \right] \log(2T/\delta)} + \sqrt{5} n \log(2T/\delta) \\
&\le \sqrt{4nT \log(2T/\delta)} + \sqrt{5} n \log(2T/\delta),
\end{aligned}
$$

where in the second inequality we use the Cauchy-Schwarz inequality. □

Proposition D.6 calculates the regret incurred from exceeding the inventory constraints. Denote $(\cdot)^+ = \max\{\cdot, 0\}$. This proposition allows us to understand the relationship between resource allocation decisions and the incurred regret due to inventory limitations.

Proposition D.6. *We have*

$$
\begin{aligned}
&\mathbb{E}\left[\underbrace{\sum_{i=1}^{n} r_i \left[\sum_{t=1}^{T} \sum_{j=1}^{L} \mathbb{I}(i,j,t) a^t - c_i \right]^+}_{(*)} \right] \\
&\le \sqrt{16 \max_{i \in n} |\Theta_i| nT \log(2T/\beta(n,T))} + \sqrt{5} n \log(2T/\beta(n,T)) + n(T \log T + 2T)\beta(n,T).
\end{aligned}
$$

Proof. We rewrite

$$
\begin{aligned}
(*) =& \sum_{i=1}^{n} r_i \Bigg[\underbrace{\sum_{t=1}^{T}\sum_{j=1}^{L} \mathbb{I}(i,j,t)(a^t - \bar{f}_i(x^{(j)}, \Omega_i^{t-1}))}_{(**)} \\
&+ \underbrace{\sum_{t=1}^{T}\sum_{j=1}^{L} (\mathbb{I}(i,j,t) - \mu_j \bar{s}_{ij}^t)\bar{f}_i(x^{(j)}, \Omega_i^{t-1})}_{(***)} + \underbrace{\sum_{t=1}^{T}\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) - c_i}_{(****)} \Bigg]^+ \\
\le& \sum_{i=1}^{n} r_i |(**)| + \sum_{i=1}^{n} r_i |(***)| + \sum_{i=1}^{n} r_i [(****)]^+.
\end{aligned}
$$

From step 3 of the algorithm and the Cauchy-Scharz inequality, we know given $\mathcal{E}$,

$$
\begin{aligned}
\sum_{i=1}^{n} |(**)| =& \sum_{i=1}^{n} \left| \sum_{\theta_i \in \Theta_i} \sum_{t' \in D_i^T(\theta_i)} \mathbb{I}(i, J^t, t') \left(a^{t'} - \bar{f}_i(x^{(J^{t'})}, \bar{\theta}^{t'}) \right) \right| \\
\le& \sum_{i=1}^{n} \sum_{\theta_i \in \Theta_i} \sqrt{2|D_i^T(\theta_i)| \log(2T/\beta(n,T))} \\
\le& \sqrt{2n \max_{i \in [n]} |\Theta_i| T \log(2T/\beta(n,T))}
\end{aligned}
$$

holds almost surely. From Lemma D.5 we know

$$
\mathbb{P}\left[\sum_{i=1}^{n} |(***)| \le \sqrt{4nT\log(2T/\beta(n,T))} + \sqrt{5}n\log(2T/\beta(n,T)) \right] \ge 1 - \beta(n,T).
$$

Recall $\bar{s}_{ij}^t$ is the optimal solution of LP (2). Thus, noticing $\lambda_j = T\mu_j$, it holds for all $i \in [n]$ almost surely

$$
\sum_{j=1}^{L} \mu_j \bar{s}_{ij}^t \bar{f}_i(x^{(j)}, \Omega_i^{t-1}) \le c_i/T.
$$

Hence, we know $(****) \le 0$holds almost surely.

Therefore, by setting $\alpha = \max_{i\in[n]} r_i \sqrt{16 \max_{i\in[n]} |\Theta_i| nT \log(2T/\beta(n,T))} + \sqrt{5}n\log(2T/\beta(n,T))$, we obtain

$$
\mathbb{P}[(*) \le \alpha | \mathcal{E}] \ge 1 - \beta(n,T).
$$

Thus, noticing

$$
\begin{aligned}
\mathbb{P}[((*) \le \alpha, \mathcal{E})^c] =& 1 - \mathbb{P}[(*) \le \alpha, \mathcal{E}] \\
=& 1 - \mathbb{P}[(*) \le \alpha | \mathcal{E}]\mathbb{P}[\mathcal{E}] \\
\le& 1 - (1 - \beta(n,T))(1 - (\log T + 1)\beta(n,T)) \\
\le& (\log T + 2)\beta(n,T),
\end{aligned}
$$

we then have

$$
\begin{aligned}
\mathbb{E}[(*)] =& \mathbb{E}[(*) | (*) \le \alpha, \mathcal{E}]\mathbb{P}[(*) \le \alpha, \mathcal{E}] + \mathbb{E}[(*) | ((*) \le \alpha, \mathcal{E})^c]\mathbb{P}[((*) \le \alpha, \mathcal{E})^c] \\
\le& \alpha + (\log T + 2)\beta(n,T)nT \\
=& \max_{i\in[n]} r_i \sqrt{16 \max_{i\in[n]} |\Theta_i| nT \log(2T/\beta(n,T))} + \sqrt{5}n\log(2T/\beta(n,T)) + n(T\log T + 2T)\beta(n,T).
\end{aligned}
$$

□

Theorem D.7 provides a summary of the expected regret bound for the $\mathrm{ALG_{LP}}$ algorithm. Specifically, the regret bound, denoted by $\tilde{O}\left(\sqrt{\max i \in [n]|\Theta_i|nT}\right)$, indicates that the expected regret increases with the square root of the maximum parameter space size $|\Theta_i|$, the number of resources n, and the time horizon T. This regret bound is particularly relevant in scenarios where customer arrivals exhibit a near-i.i.d. pattern, and it allows us to assess the scalability and performance of the $\mathrm{ALG_{LP}}$ in different settings and make informed decisions regarding resource allocation.

Theorem D.7. *The expected regret of the algorithm is upper bounded by*

$$8\max_{i\in[n]} r_i\sqrt{\max_{i\in[n]}|\Theta_i|nT\log(2nT^2)} + \sqrt{5}n\log(2nT^2) + \bar{U}/(nT^2) + LP(\theta^*)(\log T+1)/(nT) + \log T + 2.$$

Proof. The expected regret is upper bounded by the sum of the regret incurred by the expected revenue and violation of inventory constraints, that is,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} r_i \sum_{j=1}^{L}\mu_j(s^*_{ij}-\bar{s}^t_{ij})f_i(x^{(j)},\theta^*_i)\right] + \mathbb{E}\left[\sum_{i=1}^{n} r_i\left[\sum_{t=1}^{T}\sum_{j=1}^{L}\mathbb{I}(i,j,t)a^t - c_i\right]^+\right].$$

So by piecing together Proposition D.4 and D.6 we obtain the result. $\square$

Moreover, a key feature of our regret bound is that it is independent of the number of contexts L, which means it might lead to efficient computation even when L is very large, as long as we have a moderate $\max_{i\in[n]}|\Theta_i|$. This feature is also shared by (Badanidiyuru et al., 2014), where they have the cardinality of the policy space instead of the contextual space appear as a multiplicative factor in the regret bound (this allows them to treat resourceful contextual bandit problems with an infinite-dimensional contextual space).

Remark D.8. When $\min_{i\in[n]} c_i$ is very large, i.e., tends to infinity, it is easy to verify

$$(1+\min_{i\in[n]} c_i)\left(1-e^{-1/\min_{i\in[n]} c_i}\right) \to 1$$

, i.e., the CR has a limit $1+\frac{1}{1-1/e}$.

D.2. Proofs under Adversarial Arrivals

Before analyzing the regret, we introduce an auxiliary online learning problem where we maintain the adversarial arrival setting but remove the inventory constraint that limits the consumption of each resource to at most c_i units, for all $i \in [n]$. Suppose we have an algorithm Π that decides I^t in each period after observing x^t. The regret in this auxiliary problem, denoted as $\mathrm{REG_{AUX}}$, is defined as the sum of regrets incurred in each period t:

$$\mathrm{REG_{AUX}} = \sum_{t\in[T]}(r_{I^t_*}f_{I^t_*}(x^t,\theta^*_{I^t_*}) - r_{I^t}f_{I^t}(x^t,\theta^*_{I^t})),$$

where I^t_* represents the optimal resource selection that maximizes the reward $r_i f_i(x^t,\theta^*_i)$ among all resources $i\in[n]$, denoted as $I^t_* := \arg\max_{i\in[n]} r_i f_i(x^t,\theta^*_i)$. By studying the regret in this auxiliary problem, we can gain insights into the impact of resource allocation decisions on the incurred regret, without considering the inventory constraint.

Our analysis is based on a crucial finding in (Cheung et al., 2022) (c.f., Lemma D.9), which says as long as we have an online algorithm for the auxiliary problem with low time regret, we can adapt the same algorithm to the complete problem within the same time regret plus some approximation error (unrelated to T).

Lemma D.9. *((Cheung et al., 2022)) Let Π be some algorithm that incurs regret REG_{AUX} for the auxiliary online learning problem (without inventory constraints), suppose we apply the same algorithm Π to the complete online learning problem (with inventory constraints), just with r_i in period t changed to $r^t_i = r_i\left(1-\Psi\left(\frac{N^{t-1}_i}{c_i}\right)\right)$, and denote OPT as the optimal* *revenue associated with some arrival sequence $x_1, \ldots, x_T$ and ALG the revenue gained from running Π with respect to the arrival sequence $x_1, \ldots, x_T$. Then, we have*

$$OPT \leq \frac{(1 + \min_{i \in [n]} c_i)\left(1 - e^{-1/\min_{i \in [n]} c_i}\right)}{1 - 1/e} \mathbb{E}[ALG] + \mathbb{E}[REG_{AUX}].$$

So, the analysis reduces to upper bound $\mathbb{E}[\text{REG}_{\text{AUX}}]$ properly. As shown in Proposition D.10, we derive $\mathbb{E}[\text{REG}_{\text{AUX}}] = \tilde{O}(\sqrt{nT})$.

Proposition D.10. $\mathbb{E}[REG_{AUX}] \leq \max_{i \in [n]} r_i(\sqrt{|\Theta| n} + 1)\sqrt{2T \log(2T/\beta(n,T))} + \max_{i \in [n]} r_i(1/T + \log T + 1)/n.$

Proof. Note given $\mathcal{E}$, it holds almost surely

$$\begin{aligned}
\text{REG}_{\text{AUX}} =& \sum_{t=1}^{T} (r_{I_*^t} f_{I_*^t}(x^{(J^t)}, \theta^*_{I_*^t}) - r_{I^t} f_{I^t}(x^{(J^t)}, \theta^*_{I^t})) \\
=& \sum_{t=1}^{T} (r_{I_*^t} f_{I_*^t}(x^{(J^t)}, \theta^*_{I_*^t}) - r_{I^t} f_{I^t}(x^{(J^t)}, \bar{\theta}^t) + r_{I^t} f_{I^t}(x^{(J^t)}, \bar{\theta}^t) - r_{I^t} f_{I^t}(x^{(J^t)}, \theta^*_{I^t})) \\
\leq& \sum_{t=1}^{T} (r_{I_*^t} f_{I_*^t}(x^{(J^t)}, \theta^*_{I_*^t}) - r_{I^t} f_{I^t}(x^{(J^t)}, \bar{\theta}^t) + r_{I^t} f_{I^t}(x^{(J^t)}, \bar{\theta}^t) - r_{I^t} f_{I^t}(x^{(J^t)}, \theta^*_{I^t})) \\
\leq& \sum_{t=1}^{T} (r_{I^t} f_{I^t}(x^{(J^t)}, \bar{\theta}^t) - r_{I^t} f_{I^t}(x^{(J^t)}, \theta^*_{I^t})) \\
\leq& \max_{i \in [n]} r_i \cdot \left(\underbrace{\sum_{t=1}^{T} (f_{I^t}(x^{(J^t)}, \bar{\theta}^t) - a^t)}_{(*)} + \underbrace{\sum_{t=1}^{T} (a_t - f_{I^t}(x^{(J^t)}, \theta^*_{I^t}))}_{(**)} \right),
\end{aligned}$$

where in the first inequality we use the step 3 of the algorithm (i.e., $I^t = \arg\max_{i \in [n]} r_i^t \bar{f}_i(x^{(J^t)}, \bar{\theta}^t)$), and in the second inequality we use the fact that $f_{I_*^t}(x^{(J^t)}, \bar{\theta}^t) \geq f_{I_*^t}(x^t, \theta^*_{I_*^t})$ almost surely given $\mathcal{E}$. From the step 4 of the algorithm, we know given $\mathcal{E}$, it holds almost surely

$$\begin{aligned}
(*) =& \sum_{i=1}^{n} \sum_{\theta_i \in \Theta_i} \sum_{t' \in D_i^T(\theta_i)} (f_{I^{t'}}(x^{t'}, \theta_i) - a^{t'}) \\
\leq& \sum_{i=1}^{n} \sum_{\theta_i \in \Theta_i} \sqrt{2 D_i(\theta_i) \log(2T/\beta(n,T))} \\
\leq& \sqrt{2|\Theta| n T \log(2T/\beta(n,T))},
\end{aligned}$$

where in the last inequality we use the Cauchy-Schwarz inequality. Then, using D.2 we know

$$\mathbb{P}\left[(**) > \sqrt{2T \log(2T/\beta(n,T))} \middle| \mathcal{E} \right] \leq \beta(n,T)/T.$$

Let $\alpha = \max_{i \in [n]} r_i(\sqrt{|\Theta| n} + 1)\sqrt{2T \log(2T/\beta(n,T))}$, we know

$$\mathbb{P}[\text{REG}_{\text{AUX}} \leq \alpha | \mathcal{E}] \geq 1 - \beta(n,T)/T.$$

Hence,

$$\begin{aligned}\mathbb{E}[\text{REG}_{\text{AUX}}] =&\mathbb{E}[\text{REG}_{\text{AUX}}|\text{REG}_{\text{AUX}} \leq \alpha, \mathcal{E}]\mathbb{P}[\text{REG}_{\text{AUX}} \leq \alpha, \mathcal{E}] \\ &+ \mathbb{E}[\text{REG}_{\text{AUX}}|(\text{REG}_{\text{AUX}} \leq \alpha, \mathcal{E})^c]\mathbb{P}[(\text{REG}_{\text{AUX}} \leq \alpha, \mathcal{E})^c] \\ \leq& \alpha + \max_{i\in[n]} r_i T(1/T + \log T + 1)\beta(n,T) \\ =& \max_{i\in[n]} r_i(\sqrt{|\Theta|n} + 1)\sqrt{2T\log(2T/\beta(n,T))} + \max_{i\in[n]} r_i(1/T + \log T + 1)/n.\end{aligned}$$

□

Using Lemma D.9 and Proposition D.10 we obtain the main result of this section immediately.

Theorem D.11. *It holds that*

$$\begin{aligned}OPT \leq& \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\mathbb{E}[ALG] \\ &+ \max_{i\in[n]} r_i(\sqrt{|\Theta|n} + 1)\sqrt{2T\log(2T/\beta(n,T))} + \max_{i\in[n]} r_i(1/T + \log T + 1)/n.\end{aligned}$$

Theorem D.11 provides a key result of ALG_{ADV} regret upper bound under adversarial arrivals. Specifically, it establishes that the regret incurred by the algorithm exhibits a sublinear T growth rate with a constant competitive ratio. Sublinear regret indicates that the corresponding algorithm's performance gradually approaches the performance level of the optimal offline policy as time progresses.The constant competitive ratio highlights the algorithm's performance relative to an optimal decision-making strategy. It provides a quantitative measure of the algorithm's effectiveness, indicating that it achieves regret that is at most a constant factor compared to the optimal strategy. Moreover, this result forms a crucial foundation for our subsequent analysis of the regret bound for the Unified Learning algorithm under nonstationary arrivals in Section D.3.

D.3. Proofs under Nonstationary Arrivals

D.3.1. Regret analysis: sublinear regret as long as ALG_{LP} is not switched

We first provide a proposition that states ALG_{LP} always incurs a sublinear regret and a linear rate of resource consumption with high probability as long as the switch does not happen.

Proposition D.12. *Suppose ALG_{LP} runs for t iterations without being switched, then the expected regret up to time t is upper bounded by*

$$16\sqrt{2|\Theta|nt\log(4|\Theta|t/\beta(n,T))} + \sqrt{5}n\log(2t/\beta(n,T)) + (2/t + 2 + \log t)\beta(n,T)(LP(\theta^*) + nt).$$

Moreover, each resource $i \in [n]$ has at least

$$\left[\frac{T-t}{T}c_i - \sqrt{8nt\log(2t/\beta(n,T))}\right]^+$$

remaining with probability at least $1 - \beta(n,T)$.

Proof. Given $\mathcal{E}(t)$, we know the first condition in step 3 of the algorithm ensures

$$\left|\sum_{l=1}^{t}\sum_{i=1}^{n} r_i(s^*_{iJ^l} - \bar{s}^l_{iJ^l})f_i(x^{(J^l)}, \theta^*)\right| \leq \max_{i\in[n]} r_i\sqrt{32t\log(4|\Theta|t/\beta(n,T))}.$$

And by using Lemma C.4 we derive with probability at least $1 - \beta(n,T)/t$,

$$\left|\sum_{l=1}^{t}\sum_{i=1}^{n} r_i\sum_{j=1}^{L}\mu^l_j(s^*_{ij} - \bar{s}^l_{ij})f_i(x^{(j)}, \theta^*)\right| \leq \max_{i\in[n]} r_i\sqrt{128t\log(4|\Theta|t/\beta(n,T))}.$$

The second condition in step 3 of the algorithm ensures

$$\forall\, i \in [n] : \sum_{l=1}^{t} \bar{s}_{iJ^l}^{l} \bar{f}_i(x^{(J^l)}, \Omega^{l-1}) \leq \frac{t}{T} c_i + \sqrt{2t \log(2t/\beta(n,T))}.$$

And by using C.5 we derive with probability at least $1 - \beta(n,T)/t$,

$$\forall\, i \in [n] : \sum_{l=1}^{t} \sum_{j=1}^{L} \bar{s}_{ij}^{l} \bar{f}_i(x^{(j)}, \Omega^{l-1}) \leq \frac{t}{T} c_i + \sqrt{8nt \log(2t/\beta(n,T))}.$$

Therefore, using the same logic of proving Proposition D.6, we derive with probability at least $1 - \beta(n,T)$,

$$\sum_{i=1}^{n} \left[\sum_{t=1}^{T} \sum_{j=1}^{L} \mathbb{I}(i,j,t) a^t - c_i \right]^+ \leq \sqrt{64|\Theta| nt \log(2t/\beta(n,T))} + \sqrt{5} n \log(2t/\beta(n,T)).$$

Therefore, let $\alpha := 16\sqrt{2|\Theta| nt \log(4|\Theta| t/\beta(n,T))} + \sqrt{5} n \log(2t/\beta(n,T))$we have

$$\mathbb{P}[\text{Regret}^t > \alpha | \mathcal{E}(t)] \leq (2/t + 1)\beta(n,T).$$

So we end up with

$$\begin{aligned}\mathbb{E}[\text{Regret}^t] =& \mathbb{E}[\text{Regret}^t | \text{Regret}^t \leq \alpha, \mathcal{E}(t)] \mathbb{P}[\text{Regret}^t \leq \alpha, \mathcal{E}(t)] \\ &+ \mathbb{E}[\text{Regret}^t | (\text{Regret}^t \leq \alpha, \mathcal{E}(t))^c] \mathbb{P}[(\text{Regret}^t \leq \alpha, \mathcal{E}(t))^c] \\ \leq& \alpha + (2/t + 2 + \log t)\beta(n,T)(LP(\theta^*) + nt).\end{aligned}$$

□

D.3.2. Regret analysis: sublinear regret under i.i.d. arrivals

In this section, we aim to demonstrate that under i.i.d. arrivals, the switch from ALG_{LP} to ALG_{ADV} does not occur with high probability. This can be achieved by showing that the conditions for the switch to happen are unlikely to be satisfied. Once we establish that the switch does not occur under i.i.d. arrivals, we can then apply Proposition D.12 to prove that our algorithm still achieves a sublinear expected regret in this setting.

To begin, we verify the first condition in step 3 of the algorithm is not violated for all $t \in [T]$ with high probability.

Proposition D.13. *When x^t arrives in a i.i.d. fashion, given $\mathcal{E}$, we have*

$$\begin{aligned}&\mathbb{P}\left[\forall\, t \in [T], \theta \in \Omega^{t-1} : \left| \sum_{l=1}^{t} \sum_{i=1}^{n} r_i (s_{iJ^l}(\theta) - \bar{s}_{iJ^l}^{l}) f_i(x^{(J^l)}, \theta) \right| \leq \max_{i \in [n]} r_i \sqrt{32t \log(4|\Theta| t/\beta(n,T))} \right] \\ \geq& 1 - (\log T + 1)\beta(n,T).\end{aligned}$$

Proof. Using the same logic in Proposition D.1, we have for all $t \in [T]$,

$$\sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s_{ij}(\theta) - \bar{s}_{ij}^{t}) f_i(x^{(j)}, \theta) \leq \sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j \bar{s}_{ij}^{t} (\bar{f}_i(x^{(j)}, \Omega^{t-1}) - f_i(x^{(j)}, \theta)).$$

So from the θ-removal process of the algorithm we have,

$$\begin{aligned}\sum_{l=1}^{t} \sum_{i=1}^{n} r_i \sum_{j=1}^{L} \mu_j (s_{ij}(\theta) - \bar{s}_{ij}^{l}) f_i(x^{(j)}, \theta) \leq& \sum_{l=1}^{t} \mathbb{E}[r_{I^l} (\bar{f}_{I^l}(x^{(J^l)}, \Omega^{l-1}) - f_{I^l}(x^{(J^l)}, \theta)) | \mathcal{F}_{l-1}] \\ \leq& \max_{i \in [n]} r_i \sqrt{8t \log(2t/\beta(n,T))}\end{aligned}$$

holds with probability at least $1-\beta(n,T)/(2t)$. Define probabilistic event

$$A(t) = \left\{ \forall\, \theta \in \Omega^{t-1} : \left| \sum_{l=1}^{t} \sum_{i=1}^{n} r_i (s_{iJ^l}(\theta) - \bar{s}^l_{iJ^l}) f_i(x^{(J^l)}, \theta) \right| > \max_{i \in [n]} r_i \sqrt{32t \log(4|\Theta| t/\beta(n,T))} \right\}.$$

Then using C.4 and the triangle inequality we obtain

$$\mathbb{P}[A_n] \le \beta(n,T)/t.$$

Thus by using the union bound we know the condition holds with probability at least $1-(\log T+1)\beta(n,T)$. □

We verify the second condition in step 3 of the algorithm is not violated for all $t \in [T]$ with high probability.

Proposition D.14. *When x^t arrives in a i.i.d. fashion, given $\mathcal{E}$, we have*

$$\mathbb{P}\left[\forall\, t \in [T], \forall\, i \in [n] : \sum_{l=1}^{t} \bar{s}_{iJ^l} f_i(x^{(J^l)}, \Omega^{l-1}) \le \frac{t}{T} c_i + \sqrt{2t \log(2t/\beta(n,T))} \right] \ge 1 - (\log T + 1)\beta(n,T).$$

Proof. Using C.5, it is easy to show

$$\mathbb{P}\left[\left| \sum_{l=1}^{t} \left(\bar{s}^l_{iJ^l} f_i(x^{(J^l)}, \Omega^{l-1}) - \sum_{j=1}^{L} \mu_j \bar{s}^l_{ij} f_i(x^{(j)}, \Omega^{l-1}) \right) \right| > \sqrt{2t \log(2t/\beta(n,T))} \right] \le \beta(n,T)/t.$$

Since $\bar{s}^l_{ij}$ is the solution of U^t, we have for all $i \in [n]$ it holds almost surely

$$\sum_{l=1}^{t} \sum_{j=1}^{L} \mu_j \bar{s}^l_{ij} f_i(x^{(j)}, \Omega^{l-1}) \le \frac{t}{T} c_i.$$

Therefore the desired result holds by a union bound argument. □

Finally, we obtain $\tilde{O}(|\Theta|nT)$ regret under i.i.d. arrivals, as a corollary of Proposition D.12, Proposition D.13 and Proposition D.14 .

Corollary D.15. *The expected regret under i.i.d. arrivals of the algorithm is upper bounded by*

$$16\sqrt{2|\Theta|nT \log(4|\Theta|nT^2)} + \sqrt{5}n \log(2nT^2) + (LP(\theta^*)/(nT) + 1)(2/T + 2 + 3\log T).$$

D.3.3. Regret analysis: unified regret bound where ALG_{LP} is switched to ALG_{ADV} sometime

In this section, we further analyze our algorithm to address the case where the switch occurs, transitioning from ALG_{LP} to ALG_{ADV}. We extend our analysis to nonstationary arrivals and demonstrate that our algorithm achieves sublinear regret and a constant competitive ratio. By considering the combined performance of both algorithms, we derive a unified regret bound in Proposition D.16 that provides an overall measure of our algorithm's performance. This analysis allows us to quantify the regret incurred and evaluate the effectiveness of our approach in dynamically allocating resources.

Proposition D.16. *In any case of nonstationary arrivals, the algorithm guarantees*

$$\mathit{OPT} \le \left(1 + \frac{(1 + \min_{i \in [n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1 - 1/e} \right) \mathbb{E}[\mathit{ALG}] + \tilde{O}(\sqrt{n|\Theta|T}).$$

Proof. Suppose the switch occurs in time epoch t. Denote OPT_1 the expected revenue obtained by the optimal algorithm (who knows θ^* at the beginning) from time epoch 1 to t and OPT_2 the rest of the expected revenue generated by the same algorithm. We use $\text{ALG}_{\text{LP}}^{t}$ and $\text{ALG}_{\text{ADV}}^{T-t+1}$ to denote the revenue obtained by our algorithm in the two phases, respectively. From Proposition D.12 we know

$$\text{OPT}_1 \leq \mathbb{E}[\text{ALG}_{\text{LP}}^{t}] + 16\sqrt{2|\Theta|nt\log(\frac{4t}{\beta(n,T)})} + \sqrt{5}n\log(\frac{2t}{\beta(n,T)}) + (\frac{2}{t} + 2 + \log t)\beta(n,T)(LP(\theta^*) + nt).$$

We define Empty-OPT_2 as the revenue obtained by the optimal algorithm from time $t+1$ to T, given the consumption of all resources is zero. We also define $\text{Empty-ALG}_{\text{ADV}}^{T-t+1}$ the revenue obtained by our algorithm (after the switch) in the second phase, given the consumption of all resources is zero. So from the definition, it holds almost surely

$$\text{Empty-ALG}_{\text{ADV}}^{T-t+1} \leq \text{ALG}_{\text{ADV}}^{T-t+1} + \text{ALG}_{\text{LP}}^{t}.$$

Thus, from Theorem D.11 we know

$$\begin{aligned}
\text{Empty-OPT}_2 \leq & \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\mathbb{E}[\text{Empty-ALG}_{\text{ADV}}^{T-t+1}] \\
& + \max_{i\in[n]} r_i(\sqrt{n|\Theta|}+1)\sqrt{2(T-t+1)\log(2(T-t+1)/\beta(n,T))} \\
& + \max_{i\in[n]} r_i((1/(T-t+1) + \log(T-t+1) + 1)/n.
\end{aligned}$$

For an arbitrary $0 < \alpha < 1$, we have

$$\begin{aligned}
\text{ALG}_{\text{LP}}^{t} + \text{ALG}_{\text{ADV}}^{T-t+1} &= \alpha\text{ALG}_{\text{LP}}^{t} + (1-\alpha)\text{ALG}_{\text{LP}}^{t} + \text{ALG}_{\text{ADV}}^{T-t+1} \\
&\geq \alpha\text{ALG}_{\text{LP}}^{t} + (1-\alpha)(\text{ALG}_{\text{LP}}^{t} + \text{ALG}_{\text{ADV}}^{T-t+1}).
\end{aligned}$$

Then plugging in the previous two inequalities on OPT_1, Empty-OPT_2 on the right hand side of the previous inequality, using $\text{OPT} = \text{OPT}_1 + \text{OPT}_2$, $\text{Empty-OPT}_2 \geq \text{OPT}_2$, then we have

$$\begin{aligned}
\text{OPT} &= \text{OPT}_1 + \text{OPT}_2 \\
&\leq \mathbb{E}[\text{ALG}_{\text{LP}}^{t}] + \text{Empty-OPT}_2 + \tilde{O}(\sqrt{n|\Theta|T}) \\
&\leq \mathbb{E}[\text{ALG}_{\text{LP}}^{t}] + \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\mathbb{E}[\text{Empty-ALG}] + \tilde{O}(\sqrt{n|\Theta|T}) \\
&\leq \mathbb{E}[\text{ALG}_{\text{LP}}^{t}] + \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\mathbb{E}[\text{ALG}_{\text{ADV}}^{T-t+1} + \text{ALG}_{\text{LP}}^{t}] + \tilde{O}(\sqrt{n|\Theta|T}) \\
&\leq \left(1 + \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\right)\mathbb{E}[\text{ALG}] + \tilde{O}(\sqrt{n|\Theta|T})
\end{aligned}$$

completes the proof. □

Through our analysis in Proposition D.16, we demonstrate that the ULwE algorithm designed for nonstationary arrivals achieves sublinear regret, indicating diminishing regret growth as the time horizon increases. This sublinear growth rate is a desirable property, as it reflects the algorithm's ability to effectively adapt to the changing arrival patterns and make informed resource allocation decisions. We establish that our algorithm achieves a constant competitive ratio, highlighting its performance relative to an optimal decision-making strategy. This constant competitive ratio indicates that our algorithm achieves regret which is at most a constant factor compared to the optimal strategy, demonstrating its competitiveness in resource allocation even in the face of changing customer preferences.

E. Appendix: Discussion of Imperfect Knowledge of Total Arrival Rates

In this section, we clarify our results hold more generally under the scenario where the knowledge about λ_l is approximately true. That is, now let's suppose $\lambda_l = \lambda_l^* + \delta_l$, where λ_l^* is the true arrival rate and δ_l captures some estimation error on the total arrival rate. By this definition, we regard λ_l as some quantity learned from some applied machine learning algorithm using historical data on total arrival rates. For instance, the estimated daily total arrival rate for Google.com exceeds 2 billion visits. Moreover, in Lai et al. (2016), a certain regression algorithm was employed on prediction traffic data for online advertising, resulting in a total prediction error rate of approximately 0.9%. Moreover, Cetintas et al. (2011) introduces a series of probabilistic latent class models aimed at predicting online user visits for display advertising, with experiments conducted on visit logs from tens of millions of Yahoo users, resulting in an absolute percentage error of less than 0.5. These evidences demonstrate that typically in practice, even though complete knowledge of the total arrival rates might be challenging to attain, the knowledge of approximately accurate total arrival rates is easy to obtain.

Next, we show our analysis is general enough and still holds even with this additional error term incurred in our knowledge about the total arrival rate. By the new definition of λ_l, we can carefully follow the same lines of proving Proposition D.4 to show the algorithm's expected revenue is divided into two parts: one is the expected revenue in the absence of error $(*)$, and the other is the error term introduced due to the uncertainty of λ_l $(**)$, as shown below:

$$\mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n} r_i \sum_{j=1}^{L}\mu_j(s_{ij}^* - \bar{s}_{ij}^t) f_i(x^{(j)},\theta_i^*)}_{(*)} + \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n} r_i \sum_{j=1}^{L}\delta_j(s_{ij}^* - \bar{s}_{ij}^t) f_i(x^{(j)},\theta_i^*)}_{(**)}\right]$$

We can see that due to the uncertainty of λ, an additional error term $(**)$ will arise. The function $f_i(x^{(j)}, \theta_i^*)$ denotes the probability of allocating resource i to customer type $x^{(j)}$. For each time point t and each resource i, it can only be allocated to one customer type j. So we have the magnitude of the error term $(**)$ is at most $T \max_{i,j} r_i\delta_j$. The term $T \max_{i,j} r_i\delta_j$ is thus an upper bound on the error term's impact on the expected cumulative regret in all periods. Consequently, after modifying the switching condition in ALG accordingly, the expression for overall regret in Theorem 3.1 is modified to: (let $\delta := \max_j \delta_j$)

$$\text{OPT} \leq \left(1 + \frac{(1+\min_{i\in[n]} c_i)\left(1 - e^{-1/\min_{i\in[n]} c_i}\right)}{1-1/e}\right)\mathbb{E}[\text{ALG}] + \tilde{O}(\sqrt{n|\Theta|T}) + O(T\delta)$$

From this, we conclude that as long as the magnitude of δ remains at $O(T^{-1/2})$, it will not significantly affect the overall regret. Therefore, say if we have at least T historical observations of the total arrival rates (and they are drawn from some light-tailed distributions), we can still achieve the regret order as if we have full information. Of course, we point out this is still just a pretty crude analysis and better bounds from some completely distinct algorithmic design is possible.

F. Appendix: Additional Numerical Results

F.1. Regret Table under i.i.d. Arrivals

Table 1 presents a detailed comparative analysis of the solution trajectories and regret for each algorithm under i.i.d. arrivals. It details resource allocation to Customer Types A and B and the corresponding regret values, demonstrating that ALG_{LP} and ULwE outperform ALG_{ADV}. Notably, the ALG_{LP} and ULwE algorithms maintain consistent performance without violating inventory constraints, as evidenced by the zero values in the inventory regret column. This consistency corroborates the earlier observations regarding their effective resource allocation strategies.

F.2. Regret Table under ADV Arrivals

Table 2 provides a detailed comparison of the algorithms' solution trajectories and regret. The ULwE algorithm showed a distinct shift in its allocation patterns in the later stages, aligning more closely with ALG_{ADV}. Specifically, for the initial phase

Table 1. Algorithm Solution Trajectories and Regret under i.i.d. Arrivals

	Customer A		Customer B		
Period	Resource 1	Resource 2	Resource 1	Resource 2	Regret
LP Algorithm & ULwE Algorithm					
100	0	66	12	22	4.6
200	0	126	38	36	12.4
300	0	174	71	55	20.3
400	0	235	100	65	28.5
500	0	291	132	77	36.6
ADV Algorithm					
100	0	66	0	34	0
200	19	107	10	64	10.7
300	46	128	35	91	30.0
400	76	159	53	112	47.7
500	99	192	76	133	64.7

($T < 200$), the ULwE algorithm's allocations closely resemble those of ALG_{LP}. In the middle phase ($200 < T < 300$), the ULwE algorithm shifts its approach to mirror the early strategies of ALG_{ADV}, favoring Resource B predominantly. In the final phase ($T > 300$), the ULwE algorithm adopts a more greedy allocation strategy, akin to the later strategies of ALG_{ADV}. This progression, also supported by Figure 3, emphasizes the ULwE algorithm's capacity to dynamically adjust its allocation strategies in response to evolving conditions in nonstationary environments.

F.3. Regret Table under General Arrivals

In this section, our objective is to assess the performance of ALG_{LP}, ALG_{ADV}, and the ULwE algorithm under general customer arrivals. To investigate the impact of general customer arrivals, we consider specific arrival rate settings. Table 3 provides a description of the total arrival rates (λ), which represent the expected number of arrivals, for different settings. Each row in the table corresponds to a specific setting and includes information about the time period and the total arrival rates for customer types A and B. In the i.i.d. setting, the total arrival rate for customer type A is $0.6T$, and for customer type B, it is $0.4T$, over the entire time period T. In the ADV1 setting, the time period is divided into two parts: $0.33T$ and $0.67T$. In the first part, the arrival rates for customer types A is $0.8T$ and B is $0.2T$. In the second part, the arrival rates are reversed. Similarly, other settings, such as ADV2, follow a similar pattern where the arrival rates for customers change periodically.

Regarding the performance of the i.i.d. and ADV1 settings, we have provided detailed explanations in the previous two subsections. In this section, our primary focus is directed toward the ADV2 settings, which introduce a heightened level of nonstationarity. We observe that ALG_{S} consistently performs well throughout the entire period. Conversely, ALG_{ADV} initially exhibits strong performance, followed by a period where its performance is similar to ALG_{LP}, and eventually surpasses ALG_{LP} in the later stages. This implies that under conditions of high nonstationarity, ALG_{LP} demonstrates the poorest performance.

Indeed, the findings indicate that the ULwE algorithm consistently performs well across different settings, highlighting its robustness in handling various scenarios. Specifically, it performs best in cases where there is a lower level of nonstationary arrival sequence, such as in the ADV1 setting. Since this algorithm's ability to switch from ALG_{LP} to ALG_{ADV} allows it to leverage the advantages of each approach, leading to superior performance. However, in scenarios with higher levels of nonstationarity, such as in the ADV2 settings, the performance of the ULwE algorithm may decrease but still be better than that in stationary settings.

On the other hand, the performance of the ADV algorithm is steady under nonstationary and the LP algorithm demonstrates good performance under stationary conditions, but it struggles when faced with changes in arrival rates and higher levels of nonstationarity. This implies that the algorithm's static resource allocation strategy may not be suitable for dynamically changing environments. The effectiveness of the ULwE algorithm in handling nonstationary arrival patterns and adapting its resource allocation strategies is evident. By dynamically switching between different algorithms, it can effectively respond to changing conditions and achieve better performance while minimizing regret. This adaptability is a valuable characteristic

Table 2. Algorithm Solution Trajectories and Regret under Adversarial Arrivals

	Customer A		Customer B		
Period	Resource 1	Resource 2	Resource 1	Resource 2	Regret
LP Algorithm					
100	0	13	77	10	20.5
200	0	35	141	24	39.5
300	0	75	193	32	55.8
400	0	120	241	39	71.6
500	0	153	250	77	88.1
ADV Algorithm					
100	0	13	0	87	0
200	14	21	21	144	18.0
300	44	31	66	159	40.1
400	79	44	106	174	61.3
500	100	53	150	197	81.8
ULwE Algorithm					
100	0	13	77	10	20.5
200	0	35	121	44	37.1
300	0	75	121	104	37.8
400	27	93	155	125	55.7
500	51	102	199	148	76.1

Table 3. Description of different settings for customer arrival rates.

Setting	Time Period t $(\times T)$	Arrival Rate λ_A $(\times T)$	Arrival Rate λ_B $(\times T)$
i.i.d.	1	0.6	0.4
ADV1	0.33	0.15	0.85
	0.67	0.4	0.6
ADV2	0.1	0.2	0.8
	0.3	0.8	0.2
	0.2	0.2	0.8
	0.1	0.4	0.6
	0.1	0.2	0.8
	0.1	0.02	0.98
	0.1	0.2	0.8

that allows the ULwE algorithm to optimize resource allocation and cope with the uncertainties introduced by nonstationary arrival patterns.