Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Robust Dynamic Assortment Optimization in the Presence of Outlier Customers

Xi Chen, Akshay Krishnamurthy, Yining Wang

To cite this article:
Xi Chen, Akshay Krishnamurthy, Yining Wang (2023) Robust Dynamic Assortment Optimization in the Presence of Outlier Customers. Operations Research

Published online in Articles in Advance 21 Aug 2023

. https://doi.org/10.1287/opre.2020.0281

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–17
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Robust Dynamic Assortment Optimization in the Presence of Outlier Customers

Xi Chen,[a,*] Akshay Krishnamurthy,[b] Yining Wang[c]

[a] Stern School of Business, New York University, New York, New York 10012; [b] Machine Learning Group, Microsoft Research New York City, New York, New York 10011; [c] Naveen Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080

*Corresponding author

Contact: xc13@stern.nyu.edu, https://orcid.org/0000-0002-9049-9452 (XC); Akshay.Krishnamurthy@microsoft.com (AK); yining.wang@utdallas.edu, https://orcid.org/0000-0001-9410-0392 (YW)

Received: May 8, 2020
Revised: October 18, 2022
Accepted: July 5, 2023
Published Online in Articles in Advance: August 21, 2023

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2020.0281

Abstract. We consider the dynamic assortment optimization problem under the multinomial logit model with unknown utility parameters. The main question investigated in this paper is model mis-specification under the ε-contamination model, which is a fundamental model in robust statistics and machine learning. In particular, throughout a selling horizon of length T, we assume that customers make purchases according to a well-specified underlying multinomial logit choice model in a $(1-\varepsilon)$-fraction of the time periods and make arbitrary purchasing decisions instead in the remaining ε-fraction of the time periods. In this model, we develop a new robust online assortment optimization policy via an active-elimination strategy. We establish both upper and lower bounds on the regret, and we show that our policy is optimal up to a logarithmic factor in T when the assortment capacity is constant. We further develop a fully adaptive policy that does not require any prior knowledge of the contamination parameter ε. In the case of the existence of a suboptimality gap between optimal and suboptimal products, we also established gap-dependent logarithmic regret upper bounds and lower bounds in both the known-ε and unknown-ε cases. Our simulation study shows that our policy outperforms the existing policies based on upper confidence bounds and Thompson sampling.

Funding: X. Chen acknowledges support from the National Science Foundation [Grant IIS-1845444].
Supplemental Material: The supplementary material is available at https://doi.org/10.1287/opre.2020.0281.

Keywords: dynamic assortment optimization • gap-dependent analysis • regret analysis • robustness • active elimination

1. Introduction

Operations problems, ranging from assortment optimization to supply chain management, are built on an underlying probabilistic model. When real-world outcomes follow this model, existing optimization techniques are able to provide accurate solutions. However, these model assumptions are only abstractions of reality and do not perfectly capture the sophisticated natural environment. In other words, these models are inherently mis-specified to a certain degree. Accordingly, model mis-specification and robust estimation have been important topics in the statistics literature (Huber and Ronchetti 2011). However, this literature primarily focuses on estimation or prediction from a given data set, which is insufficient for modern operations settings where decision making plays a vital role. Unfortunately, most decision-making policies are derived from optimization problems that explicitly rely on the probabilistic model, so they are inherently not robust to model mis-specification. Can we design robust policies for these operations problems?

This paper studies model mis-specification for an important problem in revenue management—dynamic assortment optimization—under a popular ε-contamination model (which will be introduced in the next paragraph). Assortment optimization has a wide range of applications in retailing and online advertising. Given a large number of substitutable products, the assortment optimization problem involves selecting a subset of products (also known as an assortment) to offer a customer such that the expected revenue is maximized. To model customers' choice behavior when facing a set of offered products, discrete choice models have been widely used, and one of the most popular such models is the *multinomial logit model* (*MNL*) (McFadden 1974). In dynamic assortment optimization, the customers' choice behavior (e.g., mean utilities of products in an MNL) is not known a priori and must be learned online, which is often the case in practice, as historical data are often insufficient (e.g., fast fashion sale or online advertising). More specifically, the seller offers an assortment

to each arriving customer for a finite time horizon T, observes the purchase behavior of the customer, and then, updates the utility estimate. The goal of the seller is to maximize the cumulative expected revenue over T periods. Because of its practical relevance, dynamic assortment optimization has received much attention in literature (Caro and Gallien 2007; Rusmevichientong et al. 2010; Saure and Zeevi 2013; Agrawal et al. 2017, 2019).

All of these existing works assume that each arriving customer makes her purchase according to an underlying choice model. Yet, in practice, a small fraction of customers could make "outlier" purchases. To model such outlier purchases, we adopt a natural robust model in the statistical literature—the ε-contamination model (Huber 1964), which dates back to the 1960s and is perhaps the most widely used model in robust statistics. In the general setup of the ε-contamination model, we are given n independent and identically distributed samples drawn from a distribution $(1-\varepsilon)P_\theta + \varepsilon Q$, where P_θ denotes the distribution of interest P, parameterized by θ (e.g., a Gaussian distribution with mean θ), and Q is an arbitrary contamination distribution. The parameter $\varepsilon > 0$, which is usually very small, reflects the level at which contamination occurs, so a larger ε value means more observations are contaminated. The standard objective is to identify or estimate the parameter θ of the distribution of interest in the presence of corrupted observations from Q. For the purpose of dynamic assortment optimization in the presence of outlier customers, the P_θ distribution represents the choice model for the majority of customers, which are "typical" (with θ being the parameter of an underlying MNL choice model of interest), whereas the Q distribution corresponds to choice models of "outlier" customers, and ε reflects the proportion of outlier customers. For dynamic assortment optimization, we also deviate from the standard parameter estimation objective and focus on designing online decision-making policies.

In the classical ε-contamination model, the "outlier distribution" Q stays stationary for all samples. To make the contamination model more practical in the online assortment optimization setting, we strengthen the model in two ways.

1. Instead of assuming a fixed corruption distribution Q for all outlier customers, we allow Q to change over different time periods (i.e., Q_t is the outlier distribution for customers at time period t).

2. Instead of assuming that each time t is corrupted "uniformly at random," we assume that outlier customers appear in at most εT time periods. The purchase pattern and arrivals of outlier customers can, however, be arbitrary and even *adaptive* to the assortment decisions or customer purchase activities prior to time period t. The corrupted time periods and associated Q_t's are unknown to the seller.

This setting is much richer than the "random arrival setting" and more realistic in practice. Indeed, in a holiday season, consecutive time periods might contain anomalous or outlier purchasing behavior, which cannot be capture by "random corruption" in the original ε-contamination model. The details of our outlier customer model will be rigorously specified in Section 3.

The main goal of the paper is to develop a robust dynamic assortment policy under this ε-contaminated MNL. Our first observation is that popular policies in the literature, including upper confidence bounds (UCBs) (Agrawal et al. 2019) and Thompson sampling (TS) (Agrawal et al. 2017), no longer work in this model. The reason is that these policies cannot use typical customers who arrive later in the selling period to correct for misleading customers who arrive early on, and hence, even a small number of outlier customers can lead to poor performance. Further, although it is well known that *randomization* is crucial in any adversarial setting (see, e.g., Auer et al. 2002, Bubeck and Cesa-Bianchi 2012) to hedge against outliers, UCB is a deterministic policy, whereas Thompson sampling provides very little randomization via posterior sampling. We explain these failures in more detail in Sections 3 and 8 later in this paper.

To address the contaminated setting, we develop a novel active-elimination algorithm for robust dynamic planning, which gradually eliminates those items that are not in the optimal assortment with high probability (see Algorithm 1). Compared with the existing methods mentioned (Agrawal et al. 2017, 2019), our active-elimination method has several important technical novelties. First, our active-elimination policy implements the randomization in a much more explicit way by sampling from a carefully constructed small set of "active" products. Second, the existing UCB and Thompson sampling algorithms for MNL rely on an epoch-based strategy (i.e., repeatedly offering the same assortment until no purchase) to enable an unbiased estimation of utility parameters. This procedure is inherently fragile because the stopping time of an epoch relies on a single no-purchase activity, which can be easily manipulated by outlier customers; a few outliers can greatly affect the stopping times. The failure of such an epoch-based strategy implies that unbiased estimation of utility parameters is no longer possible. To overcome this challenge, we propose a new utility estimation strategy based on geometrically increasing offering time periods. We conduct a careful perturbation analysis to control the bias of these estimates, which leads to new confidence bounds for our active-elimination algorithm (see Section 4 for more details).

We provide theoretical guarantees for our proposed robust policy via regret analysis and information-theoretic lower bounds. In particular, let T be the selling horizon, N be the total number of products, and K be

the cardinality constraint of an assortment (see Section 3). For the reasonable setting where ε is not too large, our active-elimination algorithm (Algorithm 1) achieves $\widetilde{O}(\varepsilon K^2 T + \sqrt{KNT})$ regret when ε (or a reasonable upper bound of ε) is known (see Theorem 1), where $\widetilde{O}(\cdot)$ only suppresses $\log(T)$ factors. Compared with the $\Omega(\varepsilon T + \sqrt{NT})$ lower bound (see Proposition 1), our upper bound is tight up to polynomial factors involving K and other logarithmic factors. We also remark that the special case of $\varepsilon = 0$ reduces to the existing setting studied in Agrawal et al. (2017, 2019) and Chen and Wang (2018), in which no outlier customers are present. Compared with existing results, our regret bound is tight except for an additional $O(\sqrt{K})$ factor, which represents the cost of being adaptive to outlier customers (see Section 4.2 for more discussions). We emphasize that in a typical assortment optimization problem, the capacity of an assortment K is usually a small constant, especially relative to T and N.

The result assumes that an upper bound on the outlier proportion ε is given as prior knowledge. Although in some cases, we may be able to estimate ε from historical data, this is not always possible, which motivates the design of fully adaptive policies that do not require ε as an input. Inspired by the "multilayer active arm race" from the multiarmed bandits (MABs) literature (Lykouris et al. 2018), we propose an adaptive robust dynamic assortment optimization policy in Algorithm 3. Our policy runs multiple "threads" of known-ε algorithms on a geometric grid of ε values in parallel, and as we show, it achieves $\widetilde{O}(\varepsilon T + \sqrt{NT})$ regret, where $\widetilde{O}$ suppresses $\log(T)$ and K factors (see Theorem 2). Here, the (cumulative) regret is defined as the sum of the differences between the expected rewards (revenues) of the optimal assortment and the assortments the retailer offers at each time period. Algorithm 3 and its analysis in Section 5 provide more details.

Finally, in the case of well-separated problem instances (i.e., there is a large suboptimality gap $\beta > 0$ between optimal and suboptimal assortments), built on the same proposed algorithm, we establish much improved regret upper bounds of $\widetilde{O}(\varepsilon K^2 T \log T + K^2 N \log^2 T/\beta)$ when ε is known (see Theorem 3). When ε is unknown, the adaptive policy achieves the regret $\widetilde{O}(\varepsilon T + N/\beta^2) \times \mathrm{poly}(K, \log(NT))$ or $\widetilde{O}(\varepsilon T/\beta + N/\beta) \times \mathrm{poly}(K, \log(NT))$, whichever is smaller (see Theorem 4). For both upper bounds in the well-separated case, the dependency on the time horizon T is logarithmic when the corruption level ε is small. We also prove lower bounds on the regret when a suboptimality gap of at least $\beta > 0$ exists.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the problem formulation. The first active-elimination policy and the regret bounds are presented in Section 4, whereas the adaptive algorithm is presented in Section 5. The gap-dependent regret analysis and log T-type regret bounds are provided in Section 6. Numerical illustrations are provided in Section 8, with the conclusion in Section 9. The proof the lower-bound result is provided in the supplementary material. Proofs of some technical lemmas are relegated to the supplementary material as well.

2. Related Works

Static assortment optimization with known choice behavior has been an active research area since the seminal works by van Ryzin and Mahajan (1999) and Mahajan and van Ryzin (2001). Motivated by fast fashion retailing, dynamic assortment optimization, which adaptively learns unknown customers' choice behavior, has received increasing attention in the context of data-driven revenue management. The work by Caro and Gallien (2007) first studied the dynamic assortment optimization problem under the assumption that demands for different products are independent. Recent works by Rusmevichientong et al. (2010), Saure and Zeevi (2013), Agrawal et al. (2017, 2019), Chen and Wang (2018), and Chen et al. (2021a, b) incorporated MNL models into dynamic assortment optimization and formulated the problem as an online regret minimization problem. In particular, for the standard MNL model, Agrawal et al. (2017, 2019) developed UCB- and Thompson sampling-based approaches for online assortment optimization. Moreover, some recent works (Cheung and Simchi-Levi 2017, Oh and Iyengar 2019, Chen et al. 2020) study dynamic assortment optimization based on contextual MNL models, where the utility takes the form of an inner product between a feature vector and the coefficients. The present work focuses on the standard noncontextual MNL model, but a natural direction for future work is to extend our results to the contextual setting.

All works outlined assume that an underlying MNL choice model is correctly specified. However, model mis-specification is common in practice, and robust statistics, one of the most important branches in statistics, is a natural tool to address such mis-specification. The ε-contamination model, which was proposed by Huber (1964), is perhaps the most widely used robust model and has recently attracted much attention from the machine learning community (see, e.g., Chen et al. 2016; Diakonikolas et al. 2017, 2018; and reference therein). Despite this attention, online learning in the ε-contamination model or its generalizations is relatively unexplored. In the online setting, Esfandiari et al. (2018) studied online allocation under a mixing adversarial and stochastic model, but the setting does not require any learning component. For online learning, the recent works of Lykouris et al. (2018) and Gupta et al. (2019) studied the contaminated stochastic MAB,

but because of the complex structure of discrete choice models, these results do not directly apply to our setting. Indeed, a straightforward analogy between assortment optimization and MAB is to treat each feasible assortment as an arm, but directly using this mapping will result in a large regret because of the exponentially many possible assortments.

In learning and decision-making settings, a few recent works investigate the impact of model mis-specification in revenue management (e.g., see Cooper et al. 2006 for capacity booking problems and Besbes and Zeevi 2015 for dynamic pricing). In particular, Besbes and Zeevi (2015) show that a class of pricing policies based on linear demand functions performs well even when the underlying demand is not linear. Cooper et al. (2006) also identified some cases where simple decisions are optimal under mis-specification. However, our setting is quite different, as the widely used UCB and Thompson sampling policies are not robust under our model. On the other hand, our new active-elimination policy is robust to model mis-specification and additionally, achieves near-optimal regret when the model is well specified.

Finally, the successive-elimination and active-elimination strategies have been extensively studied in the (stochastic) multiarmed bandit literature. Interested readers can refer to the works of Auer (2002), Even-Dar et al. (2006), and Auer and Ortner (2010) for details.

3. Problem Formulation

There are N items, each associated with a known revenue parameter $r_i \in [0,1]$ and an unknown utility parameter $v_i \in [0,1]$. At each time t, a customer arrives for a total of T time periods. The retailer then provides an *assortment* $S_t \subseteq [N]$ to the customer, subject to a capacity constraint $|S_t| \le K$. The customer then chooses *at most* one item $i_t \in S_t$ to purchase, upon which the retailer collects a revenue of r_{i_t}. If the customer chooses to purchase nothing (denoted by $i_t = 0$), then the retailer collects no revenue.

At each time t, the arriving customer is assumed to be one of the following two types.

1. A typical customer makes purchases $i_t \in S_t \cup \{0\}$ according to an MNL choice model:

$$\Pr[i_t = i \mid S_t] = \frac{v_i}{v_0 + \sum_{j \in S_t} v_j}, \quad v_0 = 1. \tag{1}$$

We assume that $v_i \in [0,1]$.

2. An outlier customer makes purchases $i_t \in S_t \cup \{0\}$ according to an arbitrary unknown distribution Q_t (marginalized on $S_t \cup \{0\}$). The distribution Q_t can potentially change with t.

We note that the MNL model in Equation (1) together with the constraint that $v_i \in [0,1]$ implies that "no purchase" is the most probable (or equally probable) outcome for a *typical* customer. This assumption has been made in the operations literature (see, e.g., Agrawal et al. 2017). Such an assumption that $v_i \le 1$ for all i is, however, only for the ease of presentation, and the assumption can be easily relaxed to $v_i \le C_v$ for some known constant upper bound $C_v > 0$. With the relaxed boundedness condition, one can enlarge the constructed confidence intervals $\widehat{\Delta}_{\bar{\varepsilon}}(\tau+1)$ (see the definition in Algorithm 1) by multiplying a C_v factor, and the other parts of our analysis/algorithms remain the same.

We consider the following ε-contamination model.

A1. (Bounded adversaries.) The number of outlier customers throughout T time periods does not exceed εT, where $\varepsilon \in [0,1)$ is a problem parameter;

A2. (Adaptive adversaries.) The choice model Q_t for an outlier customer at time t can be *adversarially* and *adaptively* chosen based on the previous customers, offered assortments, and past purchasing activity.

A rigorous mathematical formulation is as follows. For any time period $t = 1, 2, \ldots, T$, let $\phi_t \in \{0,1\}$ be the indicator variable of whether customer at time t is an outlier ($\phi_t = 1$ if customer t is an outlier and 0 otherwise), $S_t \subseteq [N]$ be the assortment provided at time t, and $i_t \in S_t \cup \{0\}$ be the purchasing activity of the customer. The protocol is formally defined as follows.

Definition 1 (Definition of the Protocol). We define the following.

1. An *adaptive adversary* consists of T arbitrary measurable functions $\mathfrak{A}_1, \ldots, \mathfrak{A}_T$, where $\mathfrak{A}_t : \{\phi_\tau, Q_\tau, S_\tau, i_\tau\}_{\tau \le t-1} \longmapsto (\phi_t, Q_t)$ produces the type of the customer (typical or outlier) ϕ_t and the outlier distribution Q_t at time period t, from the filtration $\mathcal{F}_{t-1} = \{\phi_\tau, Q_\tau, S_\tau, i_\tau\}_{\tau \le t-1}$.

2. An *admissible policy* consists of T random functions $\mathfrak{P}_1, \ldots, \mathfrak{P}_T$, where $\mathfrak{P}_t : \{S_\tau, i_\tau\}_{\tau \le t-1} \longmapsto S_t$ produces a randomized assortment $S_t \subseteq [N], |S_t| \le K$ at time period t, from the filtration $\mathcal{G}_{t-1} = \{S_\tau, i_\tau\}_{\tau \le t-1}$.

3. If $\phi_t = 0$, then i_t is realized according to model (1) conditioned on S_t; otherwise, if $\phi_t = 1$, then i_t is realized according to model Q_t.

The objective of the retailer is to develop an admissible dynamic assortment optimization strategy that is competitive with a certain "benchmark" assortment. Unlike the classical setting, the definition of regret is a bit more complicated because of the presence of both typical and adversarial customers. To shed light on the subtle differences between different benchmark assortments, in this paper we consider two different types of cumulative regret, as introduced here. To simplify notations, we use P_t to denote the customer's choice model at time t. More specifically, P_t is the "typical" model in Equation (1) (denoted as $P_t = \{v\}$) if a typical customer arrives at time t, and $P_t = Q_t$ if an outlier customer arrives at time t. We use $R(S; P)$ to denote the expected revenue collected by offering assortment S if the customer's choice model is modeled by P.

1. The *typically optimal, typically evaluated* (TOTE) regret is defined as

$$\text{Regret}^{\text{TOTE}}(T) := \mathbb{E}\left[\sum_{t=1}^{T} R(S^*;\{v\}) - R(S_t;\{v\})\right], \tag{2}$$

where $S^* = \arg\max_{S\subseteq[N],\,|S|\le K} R(S;\{v\})$ is the optimal assortment *for typical customers*.

2. The *best-in-hindsight* (BIH) regret is defined as

$$\text{Regret}^{\text{BIH}}(T) := \max_{S\subseteq[N],\,|S|\le K} \mathbb{E}\left[\sum_{t=1}^{T} R(S;P_t) - R(S_t;P_t)\right]. \tag{3}$$

The TOTE regret uses the optimal assortment for typical customers S^* as the benchmark. Furthermore, the TOTE regret is always measured in the difference of expected revenue on typical customers, regardless of whether a typical customer or an outlier customer is present at time t. On the other hand, the BIH regret measures the performance differences on the actual choice model P_t of the incoming customers. In other words, it compares the performance of the dynamic assortment planning algorithm with the optimal assortment on both typical and outlier customers. The BIH regret also coincides with the "best stationary benchmark" regret considered in most fully adversarial multiarmed bandit problems.

There is an important relationship between these two definitions of regret as characterized in the following statement.

Fact 1. $\text{Regret}^{\text{BIH}}(T) \le \text{Regret}^{\text{TOTE}}(T) + \varepsilon T$.

Proof. Let S^* be the optimal assortment for typical customers and $\widetilde{S}$ be the assortment attaining the maximum in the definition of $\text{Regret}^{\text{BIH}}(T)$. Note that during time periods t that $P_t = \{v\}$, $R(\widetilde{S};\{v\}) - R(S_t;\{v\}) \le R(S^*;\{v\}) - R(S_t;\{v\})$. During time periods t that $P_t = Q_t$, we have $|(R(\widetilde{S};Q_t) - R(S_t;Q_t)) - (R(S^*;Q_t) - R(S_t;Q_t))| \le 1$ because the expected revenue of any assortment under any choice model is at most one by normalization. Because there are εT outlier time periods, we have that $\text{Regret}^{\text{BIH}}(T) \le \text{Regret}^{\text{TOTE}}(T) + \varepsilon T$. □

Fact 1 shows that the difference between the TOTE regret and the BIH regret is at most εT. Therefore, we shall focus solely on the TOTE regret in terms of the *upper bound*, which always exhibits an εT additive term in the bounds. Such an upper bound implies the same regret bound for $\text{Regret}^{\text{BIH}}(T)$, up to a term of εT. For the lower bound, we consider the BIH regret, which is standard in the literature.

4. An Active-Elimination Policy

To motivate our policy, we first briefly explain why the popular UCBs and Thompson sampling fail in the presence of outlier customers. These algorithms are designed for the uncontaminated setting where $\varepsilon = 0$, so the confidence bounds (in UCB policies) and posterior updates (in Thompson sampling policies) are designed under the assumption that *all* customers follow the same MNL model. Unfortunately, in the presence of outlier customers, the confidence intervals are too narrow, and the posterior updates are too aggressive. With these update strategies, a small number of outlier customers preferring items unpopular to typical customers could "swing" the algorithms' parameter estimates, which can lead to the belief that these unpopular items are actually popular. This subsequently leads to poor exploration of the popular items, which eventually hurts performance. As a numerical demonstration, we construct a concrete setting in Section 8, where the performance of UCB and Thompson sampling policies degrades considerably in the presence of outlier customers.

We propose an *active-elimination* policy for dynamic assortment optimization in the presence of outlier customers. A pseudocode description is given in Algorithm 1. Although Algorithm 1 requires the knowledge of ε (or an upper bound $\overline{\varepsilon}$) (see Theorem 1) as input, we emphasize that such a requirement can be completely removed by designing more complex policies, as we will show in Section 5. To highlight our main idea, we state Algorithm 1 up front as the prior knowledge of ε simplifies both the algorithm and its analysis.

Algorithm 1 (An Active-Elimination Algorithm for Robust Dynamic Assortment Optimization)

Input: time horizon T, outlier proportion $\overline{\varepsilon}$, revenue parameters $\{r_i\}$, capacity constraint K.
Output: a sequence of assortments $\{S_t\}_{t=1}^{T}$ attaining good regret.
Set $\widehat{v}^{(0)} \equiv 1$, $\widehat{\Delta}_{\overline{\varepsilon}}(0) = 1$, $\mathcal{A}^{(0)} = [N]$, $T_0 = 128(K+1)^2 N \ln T$;
for $\tau = 0, 1, 2, \ldots$ **do**
  *Compute $S_\tau^{(i)} = \arg\max_{S\subseteq\mathcal{A}^{(\tau)},\,|S|\le K,\,i\in S} R(S;\widehat{v}^{(\tau)})$ for every $i \in \mathcal{A}^{(\tau)}$;
  Compute $\gamma^{(\tau)} = \max_{i\in\mathcal{A}^{(\tau)}} R(S_\tau^{(i)};\widehat{v}^{(\tau)})$;
  Update $\mathcal{A}^{(\tau+1)} = \{i \in \mathcal{A}^{(\tau)} : R(S_\tau^{(i)};\widehat{v}^{(\tau)}) + 2\widehat{\Delta}_{\overline{\varepsilon}}(\tau) \ge \gamma^{(\tau)}\}$;
  Set $n_i = 0$ and $n_0(i) = 0$ for all $i \in \mathcal{A}^{(\tau+1)}$; set $T_\tau = 2^\tau T_0$;
  for the next T_τ time periods **do**
    Sample $i \in \mathcal{A}^{(\tau+1)}$ uniformly at random;
    Provide the assortment $S_\tau^{(i)}$ to the incoming customer and observe purchase i_t;
    Update $n_i \leftarrow n_i + \mathbf{1}\{i_t = i\}$ and $n_0(i) \leftarrow n_0(i) + \mathbf{1}\{i_t = 0\}$;

13: **end for**

14: Update estimates $\widehat{v}_i^{(\tau+1)} = \max\{1, n_i/n_0(i)\}$ for every $i \in \mathcal{A}^{(\tau+1)}$;

15: Define $\overline{\varepsilon}_\tau = \min\{1, \overline{\varepsilon}T/T_\tau\}$, $N_\tau = |\mathcal{A}^{(\tau+1)}|$, and compute error upper bound as

$$\widehat{\Delta}_{\overline{\varepsilon}}(\tau+1) = \begin{cases} 1, & T_\tau < \dfrac{\overline{\varepsilon}T}{4(K+1)}; \\ 16K(K+1)\left(\dfrac{\overline{\varepsilon}_\tau}{2} + \sqrt{\dfrac{\overline{\varepsilon}_\tau N_\tau \ln T}{T_\tau}} + \dfrac{2N_\tau \ln T}{3T_\tau}\right) & \\ \quad + 16\sqrt{\dfrac{KN_\tau \ln T}{T_\tau}}, & \text{otherwise} \end{cases}$$

16: **end for**

17: Remarks:

18: *For any set of $\{\widehat{v}\}$, $R(S;\widehat{v}) = (\sum_{i\in S} r_i\widehat{v}_i)/(1+\sum_{i\in S}\widehat{v}_i)$; the optimization can be computed efficiently. See Section 4.1 for details.

At a high level, Algorithm 1 operates in *epochs* $\tau = 0, 1, \ldots$ with geometrically increasing lengths, and it only performs item estimation or assortment updates between epochs. At any time t, the algorithm maintains an active set of items $\mathcal{A} \subseteq [N]$ consisting of all items that could potentially form a "good" assortment and estimates of parameters $\{\widehat{v}_i\}$ for all active items i in $\mathcal{A}$. For each time period t in a single epoch τ, a *random* item i is sampled from the current active item set, and a "near-optimal" assortment is built, which must contain the target item i. Once an epoch τ ends, parameter estimates of $\{\widehat{v}_i\}$ are updated, and the active set $\mathcal{A}$ is shrunk based on the updated estimates to exclude suboptimal items. We will ensure that with high probability, the optimal assortment S^* is always a subset of active sets for all epochs (see Lemma 3).

We now detail all notation used in Algorithm 1.

$\tau \in \mathbb{N}$: the indices of *epochs* whose lengths increase geometrically ($T_\tau = 2^\tau T_0$);

$\widehat{v}^{(\tau)} \in [0,1]^N$: the estimates of preference parameters (of typical customers) at epoch τ;

$\mathcal{A}^{(\tau+1)} \subseteq [N]$: the subset of active items, which are to be explored uniformly at random in epoch τ;

$\gamma^{(\tau)} \in [0,1]$ (see step 6): the estimated expected revenue of the optimal assortment calculated based on the active item subset $\mathcal{A}^{(\tau+1)}$ and current preference estimates $\widehat{v}^{(\tau)}$;

$S_\tau^{(i)} \subseteq [N]$ (see step 5): an optimal assortment computed based on $\mathcal{A}^{(\tau+1)}$ and $\widehat{v}^{(\tau)}$, which *must include* the specific item i; this assortment is used to explore and estimate the utility parameter v_i of item i;

$n_i, n_0(i) \in \mathbb{N}$ (see step 12): counters used in the estimate of v_i; note that for any supplied assortment $S_\tau^{(i)}$, we only record the number of times a customer purchases item i (accumulated by n_i) and the number of times a customer makes no purchases (accumulated by $n_0(i)$); other purchasing activities (e.g., purchases of an item $\ell \in S_\tau^{(i)}$ other than i) will not be recorded;

$\widehat{\Delta}_{\overline{\varepsilon}}(\tau+1) \in [0,1]$: length of confidence intervals used to eliminate items from $\mathcal{A}^{(\tau+1)}$; its length depends on both the epoch index τ and the prior knowledge of the outlier proportion $\overline{\varepsilon}$.

In the rest of the section, we first give a brief description of how to compute $\widehat{S}_\tau^{(i)}$ in line 5 efficiently. Then, we detail the regret upper bound of Algorithm 1 and provide the proof.

Algorithm 2 (Assortment Optimization with Additional Constraints)

1: **Input**: revenue parameters $\{r_i\}_{i=1}^n$, estimated preference parameters $\{\widehat{v}_i\}_{i=1}^n$, must-have item i, capacity constraint K, stopping accuracy δ;

2: **Output**: assortment $\widehat{S}$, $|\widehat{S}| \le K$, $i \in \widehat{S}$ that maximizes $R(\widehat{S};\widehat{v})$.

3: Initialization: $\alpha_\ell = 0$ and $\alpha_u = 1$; $\widehat{S} = \emptyset$;

4: **while** $\alpha_u - \alpha_\ell \ge \delta$ **do**

5: $\alpha_{\text{mid}} \leftarrow (\alpha_\ell + \alpha_u)/2$;

6: For each $j \ne i$, sort $\psi_j := (r_j - \alpha_{\text{mid}})\widehat{v}_j$ in descending order, and let $\Psi := \{j \ne i : \psi_j \ge 0\}$ be the subset consisting of all items other than i with nonnegative ψ_j;

7: Compute $t := \psi_i +$ the $(K-1)\psi_j$ in Ψ with the largest values;

8: If $t \ge \alpha_{\text{mid}}$, then set $\widehat{S} = \{i\} \cup \{$the $(K-1)$ items in Ψ with the largest ψ_j values$\}$ and $\alpha_\ell \leftarrow \alpha_{\text{mid}}$; else set $\alpha_u \leftarrow \alpha_{\text{mid}}$.

9: **end while**

4.1. Solving the Optimization Problem

The implementation of most steps of Algorithm 1 is straightforward, except for the computation of the assortments $S_\tau^{(i)}$, which require further algorithmic development. This computation can be formulated as the following combinatorial optimization problem:

$$\max_{|S|\le K, i\in S} R(S;\widehat{v}) = \max_{|S|\le K, i\in S} \frac{\sum_{j\in S} r_j\widehat{v}_j}{1+\sum_{j\in S}\widehat{v}_j} \tag{4}$$

for a specific $i \in [N]$. This optimization problem is similar to the classical capacity-constrained assortment optimization (see, e.g., Rusmevichientong et al. 2010), but the additional constraint $i \in S$ in (4) yields a subtle difference. For the purpose of completeness, we provide an efficient optimization method with binary search for solving Equation (4). Pseudocode is provided in Algorithm 2.

For any $\alpha \in (0,1]$, we want to check whether there exists $S \subseteq [N]$, $|S| \le K$, $i \in S$ such that $R(S;\widehat{v}) \ge \alpha$ or equivalently, $\sum_{j\in S} r_j\widehat{v}_j \ge \alpha + \alpha\sum_{j\in S}\widehat{v}_j$. Reorganizing the

terms, we only need to check whether there exists $|S| \le K$, $i \in S$ such that $\sum_{j\in S}(r_j - \alpha)\widehat{v}_j \ge \alpha$. Because $i \in S$ must hold, we only need to check whether there exists $S' \subseteq [N]\backslash\{i\}$, $|S'| \le K-1$ such that

$$(r_i - \alpha)\widehat{v}_i + \sum_{j\in S'}(r_j - \alpha)\widehat{v}_j \ge \alpha. \tag{5}$$

In order to check whether there exists such an S', we include all $j \in [N]\backslash\{i\}$ with the largest $(K-1)$ positive values of $(r_j - \alpha)\widehat{v}_j$ into the set of S' and check whether Equation (5) is satisfied. If Equation (5) holds, the current revenue value of α can be obtained, and otherwise, the current value of α cannot be obtained. We then solve the optimization problem by a standard binary search on α. We also note that $\gamma^{(\tau)}$ in line 6 is a standard static capacitated assortment optimization, which can be solved efficiently (see Rusmevichientong et al. 2010).

4.2. Regret Analysis

The following theorem is our main regret upper-bound result for Algorithm 1.

Theorem 1. *Suppose $\overline{\varepsilon} \ge \varepsilon$ and $N \le T$. Then, there exists a universal constant $C_0 < \infty$ such that for sufficiently large T, the* TOTE *regret of Algorithm 1 is upper bounded by*

$$C_0 \times \left(\overline{\varepsilon}K^2T\log T + (K^2\sqrt{\overline{\varepsilon}} + \sqrt{K})\sqrt{NT\log^3 T} + K^2N\log^2 T\right).$$

Furthermore, if $\overline{\varepsilon} \lesssim 1/K^3$ holds, then the regret upper bound can be simplified to

$$O\left(\overline{\varepsilon}K^2T\log T + \sqrt{KNT\log^3 T}\right). \tag{6}$$

Combined with Fact 1, we know that Equation (6) also serves as an upper bound for the BIH regret.

Remark 1. We note that Theorem 1 could be implied for every value of ε if there exists an adversarial bandit algorithm that achieves $\widetilde{O}(\sqrt{NT})$ BIH regret under fully adversarial settings. For multiarmed bandit, such an algorithm exists (Auer et al. 2002), rendering gap-independent analysis of the ε-contamination model trivial. However, for the assortment selection problem under the MNL model, the work of Han et al. (2021) shows that any algorithm must suffer a regret lower bound of $\widetilde{O}\left(\min\left\{T, \sqrt{\binom{N}{K}T}\right\}\right)$ against best-in-hindsight benchmarks in the fully adversarial setting. As the term $\binom{N}{K}$ is typically prohibitively large in practice, such a negative result shows that Theorem 1 *cannot* be obtained by simple black-box reduction to a fully adversarial problem in the assortment selection question.

To complement Theorem 1, we state the following proposition establishing some lower bounds for the different types of regret considered in this paper.

Proposition 1. *Let $c_0 > 0$ be a universal constant and π be any admissible policy. Suppose also that $K < N/4$.*

1. *The BIH regret of π on worst-case problem instances is at least $c_0 \times \sqrt{NT}$.*
2. *For $0 \le \varepsilon < 1$, suppose there are $\lfloor \varepsilon T \rfloor$ outlier customers. Then, the* TOTE *regret of π on worst-case problem instances is lower bounded by at least $c_0 \times (\varepsilon T + \sqrt{NT})$.*

The first property of Proposition 1 is proved by simply setting $\varepsilon = 0$ and using existing lower-bound results for dynamic assortment planning with no outlier customers (see, e.g., Chen and Wang 2018). The proof of the second property is achieved by considering the two terms εT and $\sqrt{NT}$ separately. The complete proof of Proposition 1 is given in the supplementary material.

The claims in Proposition 1 lead to a challenging open problem on the BIH regret upper bound when $\varepsilon \gtrsim \sqrt{N/T}$, at which time the εT term would dominate the $\sqrt{NT}$ term (see Equation (6)). In such cases, we conjecture that the optimal regret upper bounds would be $\sqrt{NT}$, implying that our current result in Theorem 1 is suboptimal when ε is very large. The question of achieving $\widetilde{O}(\sqrt{NT})$ regret upper bound for *all* ε levels requires fully adversarial bandit algorithms for dynamic assortment optimization, which is very challenging and an open question as far as we know.

An important special case of Theorem 1 is $\varepsilon = \overline{\varepsilon} = 0$, which reduces to the well-studied dynamic assortment optimization problem without outlier customers. For such settings, Agrawal et al. (2017, 2019) give algorithms with a regret upper bound of $\widetilde{O}(\sqrt{NT})$, which matches the lower bound of $\Omega(\sqrt{NT})$ given in Chen and Wang (2018) up to polylogarithmic terms. Comparing their results with Theorem 1, we observe that our result at $\varepsilon = \overline{\varepsilon} = 0$ matches the $\widetilde{O}(\sqrt{NT})$ regret bound except for an additional term of $O(\sqrt{K})$. This $O(\sqrt{K})$ factor stems from our active-elimination protocol and our technique for estimating the utility parameters, both of which are essential for handling outlier customers when $\varepsilon > 0$. We believe that removing this factor is technically quite challenging and leave it as an interesting open question. We also note that the capacity constraint K is typically a very small constant in practice, and hence, an additional $O(\sqrt{K})$ term is likely negligible.

Our regret upper bound in Theorem 1 also yields meaningful guarantees when ε is not zero. For example, with $\varepsilon = O(T^{-1/4})$, meaning that $O(T^{3/4})$ of T customers are outliers, Theorem 1 provides an $O(K^2T^{3/4}\log T)$

regret upper bound. This guarantee is nontrivial because it is sublinear in T, although it is larger than the standard $\widetilde{O}(\sqrt{NT})$ bound for the uncontaminated setting. Thus, Theorem 1 reveals the trade-off and impact of a small proportion of outlier customers on the performance of dynamic assortment optimization algorithms/systems.

4.3. Proof Sketch of Theorem 1

In this section, we sketch the proof of Theorem 1. Key lemmas and their implications are given, whereas the complete proofs of the presented lemmas are deferred to the supplementary material accompanying this paper.

We first state a lemma that upper bounds the estimation error $|\widehat{v}_i^{(\tau+1)} - v_i|$.

Lemma 1. *Suppose* $T_0 \geq 128(K+1)^2 N_\tau \ln T$ *and* $\min\{1, \varepsilon T/T_\tau\} \leq 1/4(K+2)$. *With probability* $1 - O(\tau_0 N/T^2)$, *it holds for all* τ *satisfying* $T_\tau \geq \max\{\overline{\varepsilon}, \varepsilon\}T/4(K+1)$ *and* $i \in \mathcal{A}^{(\tau+1)}$ *that* $|\widehat{v}_i^{(\tau+1)} - v_i| \leq \Delta_\varepsilon^*(i, \tau+1)$, *where*

$$\Delta_\varepsilon^*(i,\tau+1) = 8(K+1)\left(\frac{\varepsilon_\tau}{2} + \sqrt{\frac{\varepsilon_\tau N_\tau \ln T}{T_\tau}} + \frac{2N_\tau \ln T}{3T_\tau}\right) + 8\sqrt{\frac{(1+V_S)v_i N_\tau \ln T}{T_\tau}}, \tag{7}$$

where ε_τ *is defined as* $\varepsilon_\tau = \min\{1, \varepsilon T/T_\tau\}$, $N_\tau = |\mathcal{A}^{(\tau+1)}|$ *and* $V_S = \sum_{j \in S_\tau^{(i)}} v_j$.

Lemma 1 shows that, with high probability, the estimation error between $\widehat{v}_i^{(\tau+1)}$ and v_i, the true preference parameter of item i for typical customers, can be upper bounded by $\Delta_\varepsilon^*(i, \tau+1)$, which is a function of K, τ, T, ε, and $N_\tau = |\mathcal{A}^{(\tau+1)}|$. It should be noted that the definition of $\Delta_\varepsilon^*(i, \tau+1)$ involves unknown quantities (mostly $V_S = \sum_{j \in S_\tau^{(i)}} v_j$) and hence, cannot be directly used in an algorithm. The definition of $\widehat{\Delta}_{\overline{\varepsilon}}(\tau+1)$ in Algorithm 1, on the other hand, involves only known quantities and estimates. In Corollary 1, we will establish the connection between $\Delta_\varepsilon^*(i, \tau+1)$ and $\widehat{\Delta}_{\overline{\varepsilon}}(\tau+1)$.

Our next lemma derives how the estimated expected revenue $R(S; \widehat{v})$ deviates from the true value $R(S; v)$ by using upper bounds on the estimation errors between $\widehat{v}$ and v.

Lemma 2. *For any* $S \subseteq [N]$, $|S| \leq K$, *and* $\{\widehat{v}_i\}$, *it holds that*

$$|R(S;\widehat{v}) - R(S;v)| \leq \frac{2\sum_{i \in S}|\widehat{v}_i - v_i|}{1 + \sum_{i \in S} v_i}.$$

The proof uses only elementary algebra.

Combining Lemmas 1 and 2, we show that the $\widehat{\Delta}_{\overline{\varepsilon}}(\tau)$ quantities defined in our algorithm serve as valid upper bounds on the estimation error between $R(S; \widehat{v}^{(\tau)})$ and $R(S; v)$.

Corollary 1. *For every* τ *and* $|S| \leq K$, $S \subseteq \mathcal{A}^{(\tau)}$, *conditioned on the success events of Lemma 1 on epochs up to* τ, *it holds that* $|R(S; \widehat{v}^{(\tau)}) - R(S; v)| \leq \widehat{\Delta}_\varepsilon(\tau) \leq \widehat{\Delta}_{\max\{\varepsilon, \overline{\varepsilon}\}}(\tau)$, *where* $\widehat{\Delta}$ *is defined in Algorithm 1.*

Our next lemma is an important structural lemma, which states that, with high probability, any item in the optimal assortment S^* is never excluded from active item sets $\mathcal{A}^{(\tau+1)}$ for all epochs τ.

Lemma 3. *If* $\overline{\varepsilon} \geq \varepsilon$, *then with probability* $1 - O(\tau_0 N/T^2)$, *it holds that* $S^* \subseteq \mathcal{A}^{(\tau)}$ *for all* τ.

This structural lemma yields two important consequences. First, because "good" items remain within the active item subsets $\mathcal{A}^{(\tau+1)}$, each of the assortments $S_\tau^{(i)}$ computed in step 5 of Algorithm 1 will have relatively high expected revenue. Second, the fact that $S^* \subseteq \mathcal{A}^{(\tau+1)}$ implies that the optimistic estimates $\gamma^{(\tau)}$ will always be based on the expected revenue of the actual optimal assortment $R(S^*; v)$. This justifies the elimination step 7, in which we discard all items whose best assortment has significantly lower revenue than $\gamma^{(\tau)}$.

The proof of Lemma 3 is based on an inductive argument, which shows that if S^* belongs to $\mathcal{A}^{(\tau)}$ at the beginning of every epoch τ, then any item in S^* will not be removed (with high probability) by step 7. The intuition for this is that the optimal assortment containing any $i \in S^*$ is S^* itself, whose revenue cannot be too far away from $\gamma^{(\tau)}$ because of Lemmas 1 and 2. The complete proof of Lemma 3 is provided in the supplementary material.

Finally, our last technical lemma upper bounds the per-period regret incurred by Algorithm 1.

Lemma 4. *Suppose* $S^* \subseteq \mathcal{A}^{(\tau)}$ *holds for all* τ. *Then, with probability* $1 - O(\tau_0 N/T^2)$, *for every* $\tau \leq \tau_0$ *and* $i \in \mathcal{A}^{(\tau+1)}$, *it holds that* $R(S^*; v) - R(S_\tau^{(i)}; v) \leq 4\widehat{\Delta}_{\overline{\varepsilon}}(\tau)$.

Given the established technical lemmas, we are now ready to give the proof of Theorem 1.

Proof. Let τ^* be the smallest integer such that $T_{\tau^*} \geq \overline{\varepsilon} T/4(K+1)$. For all epochs $\tau < \tau^*$, the induced cumulative regret can be upper bounded by

$$\sum_{\tau < \tau^*} T_\tau \leq T_{\tau^*} \leq \overline{\varepsilon} T. \tag{8}$$

In the rest of this proof, we upper bound the regret incurred from epochs $\tau \geq \tau^*$. By Lemma 4, the regret incurred by a single time period in epoch τ is upper

bounded by $4\widehat{\Delta}_{\overline{\varepsilon}}(\tau)$ with high probability. The total regret accumulated in epoch τ is then upper bounded by $4\widehat{\Delta}_{\overline{\varepsilon}}(\tau)\times T_\tau$. Hence, the regret accumulated on the entire T time periods is upper bounded by

$$\begin{aligned}
&\sum_{\tau=0}^{\tau_0} 4\widehat{\Delta}_{\overline{\varepsilon}}(\tau)T_\tau \\
&\lesssim \sum_{\tau=0}^{\tau_0}\left(K^2\overline{\varepsilon}_\tau + K^2\sqrt{\frac{\overline{\varepsilon}_\tau|\mathcal{A}^{(\tau+1)}|\log T}{T_\tau}} + \frac{K^2|\mathcal{A}^{(\tau+1)}|\log T}{T_\tau}\right. \\
&\quad \left. + \sqrt{\frac{K|\mathcal{A}^{(\tau+1)}|\log T}{T_\tau}}\right)\times T_\tau \\
&\le \sum_{\tau=0}^{\tau_0}\left(\frac{K^2\overline{\varepsilon}T}{T_\tau} + K^2\sqrt{\frac{\overline{\varepsilon}|\mathcal{A}^{(\tau+1)}|T\log T}{T_\tau^2}} + \frac{K^2|\mathcal{A}^{(\tau+1)}|\log T}{T_\tau}\right. \\
&\quad \left. + \sqrt{\frac{K|\mathcal{A}^{(\tau+1)}|\log T}{T_\tau}}\right)\times T_\tau \qquad (9) \\
&\le \tau_0 K^2\overline{\varepsilon}T + K^2\sqrt{\overline{\varepsilon}T\log T}\left(\sum_{\tau\le\tau_0}\sqrt{|\mathcal{A}^{(\tau+1)}|}\right) \\
&\quad + \sqrt{K\log T}\left(\sum_{\tau\le\tau_0}\sqrt{T_\tau|\mathcal{A}^{(\tau+1)}|}\right) \\
&\quad + K^2\log T\left(\sum_{\tau\le\tau_0}|\mathcal{A}^{(\tau+1)}|\right) \\
&\le \tau_0 K^2\overline{\varepsilon}T + \tau_0 K^2\sqrt{\overline{\varepsilon}NT\log T} + \tau_0 K^2 N\log T \\
&\quad + \sqrt{K\log T}\times\sqrt{\sum_{\tau\le\tau_0}|\mathcal{A}^{(\tau+1)}|}\times\sqrt{\sum_{\tau\le\tau_0}T_\tau} \qquad (10) \\
&\le K^2\overline{\varepsilon}T\log T + K^2\sqrt{\overline{\varepsilon}NT\log^3 T} + \sqrt{K\log T}\times\sqrt{\tau_0 N} \\
&\quad \times\sqrt{T} + K^2 N\log^2 T \\
&\lesssim \overline{\varepsilon}K^2T\log T + (K^2\sqrt{\overline{\varepsilon}} + \sqrt{K})\sqrt{NT\log^3 T} + K^2N\log^2 T. \qquad (11)
\end{aligned}$$

Here, in Equation (10), we apply the Cauchy–Schwartz inequality. The final inequality holds because $\tau_0 = O(\log T)$. □

5. Adaptation to Unknown Outlier Proportion ε

In this section, we describe a more complex algorithm for robust dynamic assortment optimization where the outlier proportion ε is *unknown* a priori. Inspired by the "multilayer active arm race" for multiarmed bandits, because of Lykouris et al. (2018), Algorithm 3 runs multiple "threads" of known-ε algorithms on a geometric grid of ε values in parallel while carefully coordinating between the threads. The pseudocode of the proposed adaptive algorithm is given in Algorithm 3.

We note that for two threads $j' < j$, we have $\widehat{\varepsilon}_{j'} > \widehat{\varepsilon}_j$, which implies that the confidence interval length $\widehat{\Delta}_{\hat{\varepsilon}_{j'}}(\tau+1)$ is typically longer than $\widehat{\Delta}_{\hat{\varepsilon}_j}(\tau+1)$. Therefore, the thread j' is less aggressive than the thread j in terms of eliminating items (i.e., an item eliminated by thread j may remain active in thread j'). More detailed explanations of key steps in Algorithm 3 are summarized.

Algorithm 3 (Dynamic Assortment Optimization Robust to Unknown Outlier Proportion ε)

Input: lower bound on outlier proportion $\underline{\varepsilon} = 2^{-J}$, $J = \lfloor \log_2\sqrt{N/T}\rfloor + 1$;
Output: a sequence of assortments $\{S_t\}_t$ attaining good regret for any ε;
Construct a grid of outlier proportion values $\{\widehat{\varepsilon}_j\}_{j=0}^{J-1}$, where $\widehat{\varepsilon}_j = 2^{-j}$;
Construct J threads $j<J$, each with $\widehat{\varepsilon}_j$ outlier proportion;
For each $i\in[N]$ and $j<J$, set $\widehat{v}^{(0),j}\equiv 1$, $\widehat{\Delta}_{\hat{\varepsilon}_j}(0)=1$, $\mathcal{A}_j^{(0)} = [N]$, $T_0 = 64(K+1)^2\ln T$;
for $\tau = 0,1,2\ldots$ **do**
  for $j = 0,1,\ldots,J-1$ **do**
    If $j>0$, then update $\mathcal{A}_j^{(\tau)} = \mathcal{A}_j^{(\tau)}\cap\mathcal{A}_{j-1}^{(\tau+1)}$;
    *Compute $\gamma_j^{(\tau)}$ and $S_{\tau,j}^{(i)}$ for each $i\in\mathcal{A}_j^{(\tau)}$, and update $\mathcal{A}_j^{(\tau+1)}$;
  end for
  for the next $T_\tau = 2^\tau T_0$ time periods **do**
    Sample thread $j<J$ with probability $\wp_j := 2^{-(J-j)}/(1-2^{-J})$;
    Sample item $i\in\mathcal{A}_j^{(\tau+1)}$ uniformly at random;
    if †there exists $\widehat{\varepsilon}_k > \widehat{\varepsilon}_j$ such that $R(\widehat{S}_{\tau,j}^{(i)};\widehat{v}^{(\tau),k}) < \gamma_k^{(\tau)} - 7\widehat{\Delta}_{\hat{\varepsilon}_k}(\tau)$ **then**
      Restart Algorithm 3 with $J\leftarrow J-1$;
    end if
    Provide assortment $S_{\tau,j}^{(i)}$ to the incoming customer, and observes purchase i_t;
    Update $n_i^j \leftarrow n_i^j + \mathbf{1}\{i_t = i\}$ and $n_0^j(i)\leftarrow n_0^j(i) + \mathbf{1}\{i_t = 0\}$;
  end for
  Update estimates $\widehat{v}_i^{(\tau+1),j} = \max\{1, n_i^j/n_0^j(i)\}$ for all $j\le J$ and $i\in\mathcal{A}_j^{(\tau+1)}$;
  For every $j\le J$, compute $\widehat{\Delta}_{\hat{\varepsilon}_j}(\tau+1)$ with T, T_τ replaced by $T_j := \wp_j T$ and $T_{\tau,j} := \wp_j T_\tau$;

22: **end for**
23: *Using the procedure outlined in Algorithm 2.
24: $^{\dagger}\widehat{v}^{(\tau),k}$ and $\gamma_k^{(\tau)}$ are estimates of v and computed $\gamma^{(\tau)}$ values maintained in thread k.

1. Independence of threads. Different threads $j < J$, which correspond to different hypothetical values of ε (denoted as $\widehat{\varepsilon}_j$), are largely independent from each other, maintaining their own parameter estimates $\widehat{v}^{(\tau),j}$, active item set $\mathcal{A}_j^{(\tau+1)}$, and confidence intervals $\widehat{\Delta}_{\hat{\varepsilon}_j}(\tau+1)$. Coordination among threads only appears in two steps in Algorithm 3: step 8, which maintains a hierarchical "nested" structure of the active item sets $\mathcal{A}_j^{(\tau+1)}$ among the threads, and step 15, which provides update rules for $J \leftarrow J-1$ by comparing the obtained optimistic assortment among different threads. Further details are given in subsequent bullets.

2. Heterogeneous sampling of different threads. At each time period t when a potential customer arrives, a *random* thread $j < J$ is selected to provide assortments. The random thread, however, is not selected uniformly at random but according to a specifically designed distribution, with the probability of selecting thread j equal to $\wp_j = 2^{-(J-j)}/(1-2^{-J})$. Intuitively, such a sampling distribution "favors" the more aggressive threads with smaller hypothetical $\widehat{\varepsilon}_j$ values.

This sampling scheme is motivated by the fact that threads with larger $\widehat{\varepsilon}_j$ values typically incur large regret because their elimination rules are conservative, so many suboptimal items i remain active for many rounds. The probability of choosing these threads with large $\widehat{\varepsilon}_j$ values should be small to ensure low regret of the overall policy.

At the same time, threads corresponding to smaller $\widehat{\varepsilon}_j$ values might also incur large regret, as their overly aggressive-elimination rule might remove the optimal assortment S^* from consideration. To avoid large regret from these threads, step 15 coordinates among all of the threads and checks for inconsistencies, as we describe here.

3. Coordination and interaction among threads. As we mentioned, the coordination and interaction among different threads only happen in steps 8 and 15 in Algorithm 3. Here, we discuss these two steps in detail.

Step 8 aims at maintaining a "nested" structure among the active subsets $\mathcal{A}_j^{(\tau+1)}$, such that $\mathcal{A}_j^{(\tau+1)} \subseteq \mathcal{A}_{j'}^{(\tau+1)}$ for any $j' \le j$ at any epoch τ. We remark that such a nested structure should be expected even without this step because thread $j' \le j$ is less aggressive than thread j, in the sense that confidence intervals $\widehat{\Delta}_{\hat{\varepsilon}_{j'}}(\tau+1)$ are typically longer than $\widehat{\Delta}_{\hat{\varepsilon}_j}(\tau+1)$. Hence, one should expect that thread j' has a larger active set. Nevertheless, because of stochastic fluctuations, such nested structures might be violated. Therefore, we explicitly enforce a nesting structure at the start of every epoch τ via step 8.

Step 15 is a statistical test that tries to detect whether $\widehat{\varepsilon}_j$ is small relative to the actual (unknown) outlier proportion ε. This test crucially ensures that we do not continue to select an overly aggressive thread, which as we have mentioned, may incur large regret because of eliminating the optimal assortment S^*. Step 15 detects such events by evaluating the optimistic assortment $S_{\tau,j}^{(\cdot)}$ using the information from threads $j' < j$, which use less aggressive-elimination rules. In detail, we check if the optimistic assortment $S_{\tau,j}^{(\cdot)}$ is near optimal using the utility estimates and confidence intervals from thread j'. If the check fails and we see that $S_{\tau,j}^{(\cdot)}$ is suboptimal, we know that thread j has eliminated the optimal assortment S^* from its active set $\mathcal{A}_j^{(\cdot)}$, which subsequently leads to the conclusion that $\widehat{\varepsilon}_j$ is too small. Then, we terminate the current thread and restart the algorithm with $J \leftarrow J-1$.

We also remark on the time complexity of Algorithm 3. There are $O(\log(T/N))$ values on the ε-grid. At each time period t, a thread $\widehat{\varepsilon}_j$ is chosen. Then, at most N combinatorial optimization problems are solved, and each combinatorial optimization takes $O(NK \log T)$ time. Therefore, the total time complexity of the proposed algorithm is $O(NKT \log^2 T)$.

In the rest of this section, we state our regret upper-bound result for the adaptive Algorithm 3 as well as a sketch of its proof.

5.1. Regret Analysis and Proof Sketch

We establish the following regret upper bound for Algorithm 3. We note that all the regret mentioned in this section is the TOTE regret.

Theorem 2. *Suppose Algorithm 3 is run with an initial value of* $J = \lfloor \log_2(\sqrt{N/T}) \rfloor + 1$. *Then, there exists a constant* $C_1 = \mathrm{poly}(K, \log(NT))$ *such that, for any* $\varepsilon \in [0, 1/2]$ *and sufficiently large* T, *the regret of Algorithm 3 is upper bounded by*

$$C_1 \times (\varepsilon T + \sqrt{NT}).$$

Remark 2. In the statement of Theorem 2, $C_1 = \mathrm{poly}(K, \log(NT))$ means $C_1 = (K \log(NT))^c$ for some universal constant $c < \infty$. For notational simplicity, we did not work out the exact constant c in the expression of C_1.

The complete proof of Theorem 2 as well as the proofs of technical lemmas are relegated to the supplementary material. Here, we sketch the key steps in the proof. The first step is the following lemma, which shows that for threads with $\widehat{\varepsilon}_j \ge \varepsilon$, the optimal

assortment S^* is never removed from their active item sets with high probability.

Lemma 5. *With probability* $1 - O(\tau_0 NJ/T^2)$, *it holds for all* τ *and* $\widehat{\varepsilon}_j \geq \varepsilon$ *that* $S^* \subseteq \mathcal{A}_j^{(\tau)}$.

Lemma 5 is similar in spirit to the structural results established in Lemma 3 for Algorithm 1, but it is only applicable to thread j with $\widehat{\varepsilon}_j \geq \varepsilon$. The remaining threads, with $\widehat{\varepsilon}_j < \varepsilon$, are too aggressive in their elimination strategy, so we cannot guarantee that $S^* \subseteq \mathcal{A}_j^{(\tau+1)}$ for all τ. We will see how to upper bound the regret from these threads later in this section.

Our next lemma analyzes step 15 of the algorithm.

Lemma 6. *If* $\widehat{\varepsilon}_J \geq \varepsilon$, *then with probability* $1 - O(\tau_0 NJ/T)$, *Algorithm 3 will not be restarted*.

At a high level, Lemma 6 states that if step 15 is triggered (which causes $J \leftarrow J - 1$ and a restart of the entire algorithm), the smallest hypothetical value $\widehat{\varepsilon}_J$ must be below the actual value of ε. First, this ensures that the algorithm does not restart too often, but more importantly, it guarantees that the actual ε always falls between $\widehat{\varepsilon}_0$ and $\widehat{\varepsilon}_J$ throughout the entire selling period.

The proof of Lemma 6 is based on Lemma 5. In particular, the condition in step 15 of Algorithm 3 compares the optimistic assortments $S_{\tau,j}^{(i)}$ in thread j with estimates in threads $j' < j$, which have larger $\widehat{\varepsilon}_j$ values. If hypothetically, $\widehat{\varepsilon}_j$ is larger than or equal to ε, then by Lemma 6, we know that $S^* \subseteq \mathcal{A}_{j'}^{(\tau+1)}$ for all $j' \leq j$, and therefore, the estimated optimality of $S_{\tau,j}^{(i)}$ should be consistent in all threads $j' \leq j$. Hence, any inconsistency detected by step 15 must imply that $\widehat{\varepsilon}_j < \varepsilon$, which justifies decreasing J.

We now present two lemmas that upper bound the regret accumulated by different threads, which requires some new notation. For $0 \leq j < J$, let $\mathsf{R}(\widehat{\varepsilon}_j)$ denote the cumulative regret incurred during the time periods in which thread j is run. Clearly, the total regret incurred is upper bounded by $\sum_{j<J} \mathsf{R}(\widehat{\varepsilon}_j)$. Using linearity of the expectation, it then suffices to upper bound $\mathbb{E}[\mathsf{R}(\widehat{\varepsilon}_j)]$ for every $j < J$. The next two lemmas provide these upper bounds for two different scenarios. For notational simplicity, we use $\lesssim$ to hide $\mathrm{poly}(K, \log(NT))$ factors.

Lemma 7. *For all* $j < J$ *satisfying* $\widehat{\varepsilon}_j \geq \varepsilon$, $\mathbb{E}[\mathsf{R}(\widehat{\varepsilon}_j)] \lesssim \sum_{\tau \leq \tau_0} \mathbb{E}[\widehat{\Delta}_{\widehat{\varepsilon}_j}(\tau) \times \wp_j T_\tau]$.

Lemma 8. *For all* $j < J$ *satisfying* $\widehat{\varepsilon}_j < \varepsilon$ *and any* $\widehat{\varepsilon}_k > \max\{\widehat{\varepsilon}_j, \varepsilon\}$, *it holds that* $\mathbb{E}[\mathsf{R}(\widehat{\varepsilon}_j)] \lesssim \sum_{\tau \leq \tau_0} \mathbb{E}[\widehat{\Delta}_{\widehat{\varepsilon}_k}(\tau) \times \wp_j T_\tau]$.

These two lemmas upper bound the total accumulated regret of threads $0 \leq j < J$ separately for the case of $\widehat{\varepsilon}_j \geq \varepsilon$ and $\widehat{\varepsilon}_j < \varepsilon$. The case of $\widehat{\varepsilon}_j \geq \varepsilon$ is relatively straightforward to prove because $S^* \subseteq \mathcal{A}_j^{(\tau+1)}$ as shown in Lemma 5, so an argument similar to the proof of Theorem 1 applies. On the other hand, the case of $\widehat{\varepsilon}_j < \varepsilon$ is more difficult because S^* might be eliminated in these threads. For Lemma 8, which considers this case, we carefully analyze the stopping rule in step 15, essentially showing that the check in step 15 will trigger as soon as the regret per time period is too high for these threads. The complete proofs of both lemmas, as well as the complete proof of Theorem 2, are deferred to the supplementary material.

6. Instance-Dependent Analysis

Recall that S^* is the optimal assortment. For any given item i, let $S^{*,(i)} = \arg\max_{|S| \leq K, S \ni i} R(S)$ is the optimal assortment containing the specific item i. Define the suboptimality "gap" β as

$$\beta := R(S^*) - \max_{i \notin S^*} R(S^{*,(i)}). \tag{12}$$

Intuitively, the suboptimality gap defined in Equation (12) measures how "well defined" the optimal assortment S^* is in the sense that the inclusion of any *nonoptimal item* $i \notin S^*$ would result in at least a drop of β in expected revenue/reward, regardless of how other products in the assortment are selected. If a problem instance has a large suboptimality gap parameter β, it implies that the optimal assortment S^* is easier to learn (because nonoptimal products are easier to be ruled out), and therefore, smaller cumulative regret is expected.

It is also worthwhile to compare the gap parameter defined in Equation (12) with those defined in earlier works. In the work of Rusmevichientong et al. (2010), a nonparametric gap β' is defined as

$$\beta' := \frac{\min\{\min_i v_i, \min_{i \neq j} |v_i - v_j|, \min_{(i,j) \neq (s,t)} |\mathcal{J}(i,j) - \mathcal{J}(s,t)|\}}{(1 + K \max_i v_i)},$$

where $\mathcal{J}(i,j) := (r_i v_i - r_j v_j)/(v_i - v_j)$. It is clear that a strictly positive β' implies that all utility parameters $\{v_i\}$ are distinct. On the other hand, it is easy to construct problem instances with duplicate v_i parameters (indicating that some products have the same utility/popularity for incoming customers) and zero β', whereas our defined suboptimality gap β could still be strictly positive. Indeed, consider the following problem instance with $n = 3$ products and $K = 2$ capacity constraint, with $(v_1, v_2, v_3) = (0.5, 0.5, 1)$ and $(r_1, r_2, r_3) = (0.2, 0.5, 0.6)$. It is easy to verify that in this problem instance, $\beta' = 0$, whereas $\beta = 0.06 > 0$.

In the remainder of this section, we will use the concept of suboptimality gap defined in Equation (12) to

improve our regret upper bounds in Theorems 1 and 2, obtaining $\log(T)$-type gap-dependent regret bounds similar to bounds for stochastic multiarmed bandits. Both our Algorithms 1 and 3 remain unchanged, whereas the regret analysis is modified to take into consideration the β parameter.

6.1. Gap-Dependent Analysis of Algorithm 1 (Known Corruption Level)

We first consider Algorithm 1 designed for the setting in which a good upper bound $\overline{\varepsilon}$ on the true corruption level ε is known. The following lemma is the key lemma in the gap-dependent setting.

Lemma 9. *Let β be defined in Equation (12), and suppose $\beta > 0$. Then, with probability $1 - O(\tau_0 N/T^2)$, for every epoch τ satisfying*

$$T_\tau \geq \kappa_0 \times \max\left\{\frac{\overline{\varepsilon}K^2T}{\beta}, \frac{K^2\sqrt{\overline{\varepsilon}NT\log T}}{\beta}, \frac{K^2N\log T}{\beta}, \frac{KN\log T}{\beta^2}\right\} \tag{13}$$

for some universal constant $\kappa_0 > 0$, it holds that $\mathcal{A}^{(\tau+1)} = S^*$.

We note that in (13), $\overline{\varepsilon}$ is an upper-bound estimate of ε. At a high level, Lemma 9 states that if T_τ is sufficiently large, the active product set $\mathcal{A}^{(\tau)}$ only consists of the optimal assortment for typical customers S^*. Intuitively, this is because when T_τ is large, the confidence bound $\widehat{\Delta}_{\overline{\varepsilon}}(\tau)$ is much shorter. When the confidence interval cannot cover the underlying suboptimality gap β, the nonoptimal products $i \notin S^*$ will be automatically eliminated. A complete proof of Lemma 9 is given in the supplementary material.

With Lemma 9, we can prove the following theorem on gap-dependent regret upper bounds for Algorithm 1 with a known upper bound $\overline{\varepsilon}$ on ε.

Theorem 3. *Let β be defined in Equation (12) and $\beta > 0$. Assume also for simplicity that $\overline{\varepsilon} \lesssim 1/K^3$. The expected cumulative* TOTE *regret of Algorithm 1 is upper bounded by*

$$C_0' \times \left(\overline{\varepsilon}K^2T\log T + \frac{K^2N\log^2 T}{\beta}\right), \tag{14}$$

where $C_0' < \infty$ is a universal constant.

We remark that the $\log^2 T$ term in the second $\frac{K^2N\log^2 T}{\beta}$ term in the regret upper bound most likely arises from the doubling epochs $\{\mathcal{A}^{(\tau)}\}$ used in our proposed active-elimination algorithms, where the total number of epochs τ_0 could be logarithmic in T. It is an interesting open technical question to further improve the second term in (14) to be linear in $\log T$, which should be possible at least in the case of ε (or its suitable upper bound $\overline{\varepsilon}$) being known.

6.2. Gap-Dependent Analysis of Algorithm 3 (Unknown Corruption Level)

When the corruption level ε is unknown and no good estimate is available a priori, Algorithm 3 partitions the possible corruption levels into a logarithmic grid $\{\widehat{\varepsilon}_j\}_{j=0}^{J-1}$ and runs Algorithm 1 on different levels of $\widehat{\varepsilon}_j$ in parallel. To analyze its regret performance from a gap-dependent perspective, we again discuss the two cases of $\widehat{\varepsilon}_j \geq \varepsilon$ and $\widehat{\varepsilon}_j < \varepsilon$ separately.

In the case of $\widehat{\varepsilon}_j \geq \varepsilon$ (i.e., overestimating the true corruption level ε), Lemma 5 shows that with high probability, the optimal assortment S^* will not be removed from $\mathcal{A}_j^{(\tau)}$. Subsequently, Lemma 9 can be directly applied, with a union bound on the failure probability over $j < J$, $\widehat{\varepsilon}_j \geq \varepsilon$, as the following corollary.

Corollary 2. *For $j < J$ and epoch τ, recall the definitions that $T_j = \wp_j T$ and $T_{\tau,j} = \wp_j T_\tau$, where $\wp_j = 2^{-(J-j)}/(1-2^{-J})$ is the sampling probability for thread j and $T_\tau = 2^\tau T_0$ is the "normal" length epoch τ. Let τ_j^* be the smallest integer such that $T_{\tau_j^*,j}$ satisfies Equation (13) or more specifically,*

$$T_{\tau_j^*,j} \geq \kappa_0' \times \max\left\{\frac{\widehat{\varepsilon}_j K^2 T_j}{\beta}, \frac{K^2\sqrt{\widehat{\varepsilon}_j NT_j\log T}}{\beta}, \frac{K^2N\log T}{\beta}, \frac{KN\log T}{\beta^2}\right\}, \tag{15}$$

where $\kappa_0' > 0$ is a universal constant. Then, for all $\tau' \geq \tau_j^*$, $\mathcal{A}_j^{(\tau)} = S^*$.

Subsequently, Lemma 7 leads to the following corollary.

Corollary 3. *For all $j < J$ satisfying $\widehat{\varepsilon}_j \geq \varepsilon$, $\mathbb{E}[\mathsf{R}(\widehat{\varepsilon}_j)] \lesssim \mathbb{E}[\sum_{\tau \leq \tau_j^*} \widehat{\Delta}_{\widehat{\varepsilon}_j}(\tau) \times \wp_j T_\tau]$, where τ_j^* is defined in Corollary 2.*

We next consider the case of $\widehat{\varepsilon}_j < \varepsilon$. Because the constraint $\mathcal{A}_{j+1}^{(\tau)} \subseteq \mathcal{A}_j^{(\tau)}$ is enforced in Algorithm 3 all the time, we know that $\mathcal{A}_j^{(\tau)} = S^*$ implies $\mathcal{A}_{j+1}^{(\tau)} = S^*$ with probability 1. Consequently, Lemma 8 implies the following.

Corollary 4. *For all $j < J$ satisfying $\widehat{\varepsilon}_j < \varepsilon$ and any $\widehat{\varepsilon}_k > \max\{\widehat{\varepsilon}_j, \varepsilon\}$, it holds that $\mathbb{E}[\mathsf{R}(\widehat{\varepsilon}_j)] \lesssim \mathbb{E}\left[\sum_{\tau \leq \tau_k^*} \widehat{\Delta}_{\widehat{\varepsilon}_k}(\tau) \times \wp_j T_\tau\right]$, where τ_k^* is defined in Corollary 2 for thread k.*

With Corollaries 2–4 in place, we are ready to state our gap-dependent analysis for Algorithm 3 with unknown corruption level ε.

Theorem 4. *Suppose Algorithm 3 runs with an initial value of $J = \lfloor \log_2(\sqrt{N/T}) \rfloor + 1$. Suppose also that the gap*

parameter β defined in Equation (12) is strictly positive. Then, the cumulative TOTE *regret of Algorithm 3 can be upper bounded by*

$$(\varepsilon T + N/\beta^2) \times \mathrm{poly}(K, \log(NT)),$$

where in the regret upper bound, we hide polynomial dependency on K and $\log N, \log T$ terms.

Remark 3. An alternative upper bound of $(\varepsilon T/\beta + N/\beta) \times \mathrm{poly}(K, \log(NT))$ can also be proved, which could be larger or smaller than the one presented in Theorem 4 depending on the values of ε and β.

Comparing Theorem 4 with Theorem 3, we notice an additional $1/\beta$ term in either the εT or the N/β term in Theorem 3. Such a worsened dependency likely arises from the layered approach taken to address unknown ε values, which also delivered suboptimal regret guarantees (compared with when ε is known a priori) in robust multiarmed bandit problems (Lykouris et al. 2018, Gupta et al. 2019).

6.3. A Lower Bound on Gap-Dependent Regret

We complement our gap-dependent regret upper-bound results in the previous sections by stating a lower bound on gap-dependent regret in dynamic assortment optimization with outlier customers.

Theorem 5. *Let K, β be constants independent of T, satisfying $\beta \le \min\{1/16, 1/K\}$ and $K \le 2$. Suppose also that ε, N can potentially change with T and that $\beta \ge \sqrt{N/T}$, $K < N/4$. Then, for sufficiently large T, the worst-case BIH regret of any admissible policy is lower bounded by*

$$c_0' \times \left(\min\{\varepsilon T, \sqrt{\varepsilon NT}\} + \frac{N \log T}{K\beta} \right),$$

where $c_0' > 0$ is a universal constant independent of N, T, K, and β.

Remark 4. The lower-bound result in Theorem 5 assumes the algorithm has full knowledge of the corruption level ε.

Remark 5. As Theorem 5 only concerns the BIH regret, a similar lower bound for the TOTE regret can be established. More specifically, the $\Omega(\varepsilon T)$ lower bound in Proposition 1 still applies because there is no additional constraints/assumptions imposed on outlier customers. Furthermore, the $\frac{N \log T}{K\beta}$ lower bound in Theorem 5 is obtained by simply setting $\varepsilon = 0$, which applies to the TOTE-regret notion too. Hence, a lower bound of $\Omega\left(\varepsilon T + \frac{N \log T}{K\beta}\right)$ can be established for the TOTE regret in the gap-dependent setting.

Comparing Theorem 5 with Theorem 3 (our regret upper bound with knowledge of ε), we notice that the $K^2 N \log^2 T/\beta$ term matches the $N \log T/(K\beta)$ term in Theorem 5 up to polynomial dependency on K and $\log T$. As discussed in the works of Agrawal et al. (2017, 2019), in revenue management applications, the capacity constraint K is usually very small and therefore, treated as a constant. On the other hand, there is a gap between the $\varepsilon K^2 T \log T$ term in the upper bound and the $\min\{\varepsilon T, \sqrt{\varepsilon NT}\}$ term in the lower bound, particularly when ε is relatively large compared with N/T. We are at the moment unsure which one is tight. However, in order for the lower bound to be tight, it requires fully adversarial algorithms for dynamic assortment optimization, which has already been an open question as discussed before. Finally, the lower bound in Theorem 5 assumes the knowledge of the corruption level ε. The lower bound for cases when ε is unknown is significantly more complicated and could involve whether the upper bounds are tight in $\log T$ terms and the distinction between regret and pseudoregret notions (Lykouris et al. 2018), which are out of the scope of this paper.

7. Uniform Contamination Models

In this section, we study an *uniform contamination* model that is slightly weaker than the fully adaptive adversary protocol defined in Definition 1. Instead of allowing for the contaminated time periods to be adversarially selected and potentially concentrated or widely spread in any manner, in this section we impose the following additional assumption to constrain the distribution and pattern of contaminated time periods.

Definition 2 (Uniform Contamination Protocol). $\{\phi_t\}_{t=1}^T$ are independent identically distributed random variables with $\Pr[\phi_t = 1] = \varepsilon$ and $Q_t \equiv Q$ for some unknown underlying outlier demand distribution Q, where $\varepsilon \in [0, 1]$ is a parameter characterizing outlier portions.

We study the uniform contamination model for two purposes. First, it allows us to construct an information-theoretical lower bound of $\Omega\left(\min\left\{\varepsilon T, \sqrt{\binom{N}{K} T}\right\}\right)$, showing that the regret upper bounds established in previous sections are tight up to K factors when $\binom{N}{K}$ is not too small. As the uniform contamination model imposes stronger conditions, such a lower bound is also applicable to the general model studied in previous sections as well. Second, with the uniform contamination model, we designed a robust planning algorithm based on the UCB framework that improves an $O(K)$ factor in the εT term of the regret upper bound.

7.1. Lower Bound

We establish the following information-theoretical lower bound on *any* admissible assortment optimization policy for the uniform contamination model.

Theorem 6. *Fix $\varepsilon \in (0,1)$, and suppose $\binom{N}{K} \geq 2$. There exists a numerical constant $\underline{C}_K > 0$ depending only on K, such that for any admissible policy π, it holds that*

$$\begin{aligned}\mathrm{Regret}^{\mathsf{BIH}}(T) &= \max_{S\subseteq[N],\,|S|\leq K} \mathbb{E}\left[\sum_{t=1}^{T} R(S;P_t) - R(S_t;P_t)\right] \\ &\geq \underline{C}_K \times \min\left\{\varepsilon T, \sqrt{\binom{N}{K} T}\right\},\end{aligned}$$

where P_t is the typical distribution if $\phi_t = 0$ and $P_t = Q$ is the outlier distribution if $\phi_t = 1$, with $\{\phi_t\}_{t=1}^T$ being realized according to the uniform contamination protocol described in Definition 2.

The proof of Theorem 6 is presented in the supplementary material. At a high level, the proof is based on a key technical result from Han et al. (2021), which states that, for any $S \subseteq [N]$, $|S| \leq K$, there exists a distribution μ over $v \in [0,1]^N$ such that $\mathbb{E}_{v\sim\mu}[\mathbb{P}_v(\cdot|S')] \equiv \mathbb{P}_0$ for all $S' \neq S$ and $\mathbb{E}_{v\sim\mu}[\mathbb{P}_v(\cdot|S')] \equiv \mathbb{P}_1$ for $S' = S$ for two different distribution $\mathbb{P}_0, \mathbb{P}_1$ where $\mathbb{P}_v$ is the distribution under the MNL model parameterized by v. Using such a construction for the outlier distribution together with standard bandit lower-bound arguments (see, e.g., Bubeck and Cesa-Bianchi 2012), we can prove Theorem 6.

Remark 6. Together with the $\Omega(\sqrt{NT})$ regret lower bound established in Chen and Wang (2018) for $K \leq N/4$, Theorem 6 implies a regret lower bound of $\Omega\left(\min\left\{\varepsilon T + \sqrt{NT}, \sqrt{\binom{N}{K} T}\right\}\right)$ for the BIH regret. As the second term involves $\binom{N}{K}$, which is typically very large, in practice the lower bound could be simplified to $\Omega(\varepsilon T + \sqrt{NT})$, which matches our lower bound for the TOTE regret in Proposition 1 in the main text.

Remark 7. The problem setting for which the lower bound in Theorem 6 applies involves notably much stronger assumptions compared with the settings studied prior to this section and in the subsequent subsection, where we will present another upper bound. This makes the lower bound mathematically stronger. More specifically, the following differences apply.

1. In the uniform contamination model, each time period is contaminated (corrupted) in a uniform, stochastic manner; on the other hand, in the general model studied in previous sections, the contamination or corruption patterns are arbitrary and adaptively adversarial.

2. The adversarial demand models $\{Q_t\}_{t=1}^T$ constructed in Theorem 6 have two additional structures.

   a. If $\{Q_t\}_{t=1}^T$ are understood as adaptively adversarially chosen distributions, then each Q_t falls into the class of MNL demand models with adaptively adversarially chosen utility parameters $\{v_t\}_{t=1}^T \subseteq [0,1]^N$, which is weaker than the assumption that each Q_t could be *any* adversarially chosen demand model.

   b. The $\{Q_t\}_{t=1}^T$ constructed in Theorem 6 also have the structure of $Q_t \equiv Q$, with Q being a fixed demand distribution that is not necessarily an MNL model. This matches the definition of uniform contamination models in Definition 2.

7.2. Upper Bound

In this section, we adapt the MNL-bandit algorithm (Agrawal et al. 2019) designed originally for the stochastic assortment optimization problem to the uniform contamination setting by using median estimators with inflated upper confidence estimates. We then derive a regret upper bound that improves an $O(K)$ factor on the contamination-related term compared with the regret upper bound obtained in Theorem 1 for the general contamination model.

Algorithm 4 (MNL Bandit with Inflated UCBs for the Uniform Contamination Model)

1: **Input**: time horizon T, outlier proportion $\overline{\varepsilon}$, revenue parameters $\{r_i\}_{i=1}^N$, capacity K;

2: **Output**: a sequence of assortments $\{S_t\}_{t=1}^T$ attaining good regret;

3: Initialize: for each $i \in [N]$, $m_i = \rho_i = L_i = 0$, $\overline{v}_i = V_\infty$, $C_1 = 4\sqrt{3\ln(NT^2)}$, $C_2 = 384(1+2\overline{\varepsilon}K)\overline{\varepsilon}K$, $C_3 = 8\overline{\varepsilon}$, $C_4 = 16\overline{\varepsilon}^2(1+K)^2$, $\tau = 1$;

4: **while** T time periods have yet been reached **do**

5: Compute $S_\tau \leftarrow \arg\max_{S\subseteq[N],\,|S|\leq K} \frac{\sum_{i\in S} r_i \overline{v}_i}{1+\sum_{i\in S}\overline{v}_i}$;

6: Offer assortment S_τ repetitively until a no-purchase action occurs; let $n_\tau(i)$ be the number of times product i is purchased for $i \in S_\tau$;

7: Update: $m_i \leftarrow m_i + 1$, $\rho_i \leftarrow \rho_i + n_\tau(i)$, $L_i \leftarrow L_i + 1 + \sum_{j\in S_\tau} n_\tau(j)$ for all $i \in S_\tau$;

8: For each $i \in S_\tau$, compute

$$\widehat{v}_i \leftarrow \frac{\rho_i}{m_i}, \quad \overline{v}_i \leftarrow \min\left\{1, \widehat{v}_i + \frac{C_1}{\sqrt{m_i}} + \frac{C_2}{m_i} + \frac{C_3 L_i}{m_i} + C_4\right\};$$

9: **end while**

Algorithm 4 gives a pseudocode description of the proposed MNL-bandit algorithm variant with inflated UCBs. Compared with the existing MNL-bandit algorithm (Agrawal et al. 2019), the key difference is the definition of the inflated upper confidence estimates $\overline{v}_i$, which involves not only the stochastic confidence term $C_1/\sqrt{m_i}$ but also, a term $C_2 L_i$ that is related to estimation errors resulting from corrupted time periods. Note

that the "offer-until-no-purchase" strategy and the inflated confidence intervals are designed specifically for the uniform contamination model in Definition 2; for the general contamination model studied in previous sections, such strategies would not work, particularly when adversarial corruptions are concentrated together, in which case the inflated confidence intervals might fail to capture the estimation error.

The following theorem upper bounds the cumulative regret of Algorithm 4.

Theorem 7. *Suppose* $\varepsilon \le \overline{\varepsilon} \in [0, 1/2]$ *and* $\{\phi_t, Q_t\}_{t=1}^T$ *are realized according to the uniform contamination model defined in Definition 2. Then, it holds that*

$$\mathrm{Regret}^{\mathsf{TOTE}}(T) \le 2(1 + 4\overline{\varepsilon}K) \times (C_4 KT + 2C_1\sqrt{KNT} + 2(C_2 + C_3 KT)\ln T),$$

where $\mathrm{Regret}^{\mathsf{TOTE}}(T)$ *is the* TOTE *regret over* T *periods defined in Equation (2), which upper bounds the BIH regret in Equation (3) by definition.*

Remark 8. Suppose C_1, C_2, C_3, C_4 are selected as in Algorithm 4 and $\varepsilon \lesssim 1/K^{1.5}$. The regret upper bound in Theorem 7 could then be reduced to $O(\sqrt{KNT}\ln(NT^2) + \overline{\varepsilon}KT\ln T)$.

Because of space constraints, the complete proof of Theorem 7 is placed in the supplementary material. Comparing with the regret upper bound in Theorem 1, we notice an improvement of an $O(K)$ factor in the $\overline{\varepsilon}T$ term.

8. Numerical Illustration

In this brief experimental section, we provide some numerical illustrations that demonstrate the robustness of our proposed policy and the benefits over existing nonrobust approaches for dynamic assortment optimization, including TS (Agrawal et al. 2017) and UCBs (Agrawal et al. 2019). We construct the following data instance.

1. K of N items have revenue parameters $r_i \equiv 1$ and preference parameters $v_i \equiv 0$.

2. For the other $(N - K)$ items, both their revenue and preference parameters (r_i, v_i) are uniformly distributed on $[0.1, 0.2]$.

3. For the first $\lfloor \varepsilon T \rfloor$ time periods, the arriving customers are outliers with choice models $Q_t \equiv Q$, where Q is an MNL-parameterized choice model with preference parameters set as $v_i' = 1$ if $v_i = 0$ and $v_i' = v_i$ otherwise.

This instance reflects two important properties of outlier customers in practice, namely that they have significantly different preferences from typical customers and that they arrive in consecutive time periods (e.g., during a holiday season). In particular, the instance consists of K items with very high revenue but very low preference parameters so that few customers will buy them. Under normal circumstances, a dynamic assortment optimization algorithm would identify the unpopularity of these K items very quickly and stop recommending them. However, as the outlier customers prefer these K items over the others, these items appear popular and profitable in the early time periods, which may mislead the algorithm. As these algorithms are highly unpopular in the latter time periods, a robust algorithm should not be severely impacted by these outlier customers.

For the baseline methods, the TS method is tuning free with a noninformative Beta(1,1) prior on each item. For the UCB algorithm, we find the value in the multiplier (C_1) when constructing upper confidence bands that gives the best performance (in the original paper of Agrawal et al. 2019, $C_1 = 48$ for theoretical purposes). Each method is run for 100 independent trials, and the mean average regret (i.e., the cumulative regret over T) is reported. The standard deviations of all the methods are sufficiently small and thus, omitted for better visualization.

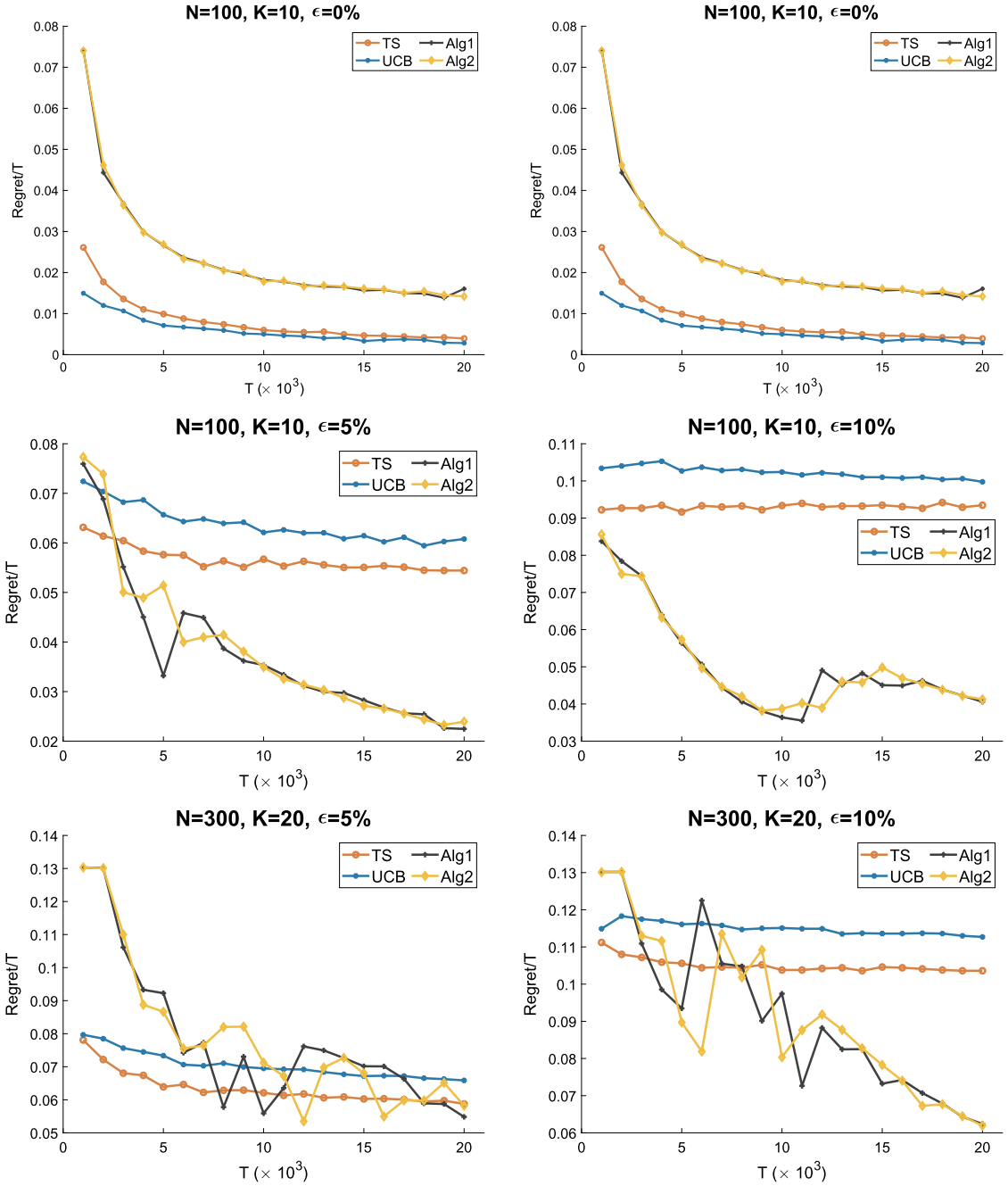
In Figure 1, we report the results for all methods under various settings of T, N, K, and ε. The experimental settings are chosen as $N \in \{100, 300\}$, $K \in \{10, 20\}$, $\varepsilon \in \{0, 0.05, 0.1\}$, and T ranging from $T = 1{,}000$ to $T = 20{,}000$. From Figure 1, we can see that when ε is strictly greater than zero, our proposed algorithms will stabilize at a mean regret level (0.02–0.06) that is much lower than the nonrobust TS and UCB methods. More importantly, the average regret (i.e., cumulative regret divided by T) for our method decreases as a function of the time horizon, a phenomenon that does not happen for TS/UCB, especially when ε is large. This confirms that these latter two methods are *not* robust to outlier customers and further confirms the effectiveness of our proposed algorithms for robust dynamic assortment optimization. For the no contamination case of $\varepsilon = 0$, whereas our proposed algorithms perform slightly worse than the baselines, the decreasing rates of average regrets are the same. When there is no contamination, although the main term in our regret $\sqrt{NT}$ is still tight, there might be extra overhead in the regret bound through dependency on K and $\log T$ factors.

9. Conclusions and Future Work

In this paper, we extend the ε-contamination model from statistics to the online decision-making setting and study the dynamic assortment optimization problem with outlier customers. We propose a new active-elimination policy that is robust to adversarial corruptions and establish a near-optimal regret bound. We further develop an adaptive policy that does not require any prior knowledge of the corruption proportion ε.

One interesting problem is to sharpen upper and lower regret bounds in the gap-dependent case. Beyond this technical question, we hope that this work

Figure 1. (Color online) Comparison of Average Regret (i.e., Regret Divided by T) Between Our Proposed Algorithms and Baselines



Note. The time horizon T ranges from 1,000 to 20,000.

inspires future work on model mis-specification in revenue management, which we believe is a practically important research direction. We look forward to pursuing this direction in future work.

Acknowledgments

The authors thank the department editor, the associated editor, and the anonymous referees for many useful suggestions and feedback, which greatly improved the paper.

References

Agrawal S, Avandhanula V, Goyal V, Zeevi A (2017) Thompson sampling for MNL-bandit. *Proc. Conf. Learn. Theory (COLT)*.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: A dynamic learning approach to assortment selection. *Oper. Res.* 67(5):1453–1485.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* 3(November):397–422.

Auer P, Ortner R (2010) UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Math. Hungarica* 61(1–2):55–65.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002) The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1):48–77.

Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Sci.* 61(4):723–739.

Bubeck S, Cesa-Bianchi N (2012) *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems* (Now Foundations and Trends, Boston).

Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.

Chen M, Gao C, Ren Z (2016) A general decision theory for Huber's ϵ-contamination model. *Electronic J. Statist.* 10(2):3752–3774.

Chen X, Wang Y (2018) A note on tight lower bound for MNL-bandit assortment selection models. *Oper. Res. Lett.* 46(5):534–537.

Chen X, Wang Y, Zhou Y (2020) Dynamic assortment optimization with changing contextual information. *J. Machine Learn. Res.* 21(216):1–44.

Chen X, Wang Y, Zhou Y (2021a) Optimal policy for dynamic assortment planning under multinomial logit models. *Math. Oper. Res.* 46(4):1639–1657.

Chen X, Shi C, Wang Y, Zhou Y (2021b) Dynamic assortment selection under nested logit models. *Production Oper. Management* 30(1):85–102.

Cheung WC, Simchi-Levi D (2017) Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. Preprint, submitted November 21, http://dx.doi.org/10.2139/ssrn.3075658.

Cooper WL, de Mello TH, Kleywegt AJ (2006) Models of the spiral-down effect in revenue management. *Oper. Res.* 54(5):968–987.

Diakonikolas I, Kamath G, Kane D, Li J, Moitra A, Stewart A (2018) Robustly learning a gaussian: Getting optimal error, efficiently. *Proc. ACM-SIAM Sympos. Discrete Algorithms*.

Diakonikolas I, Kamath G, Kane DM, Li J, Moitra A, Stewart A (2017) Being robust (in high dimensions) can be practical. *Proc. Internat. Conf. Machine Learn*.

Esfandiari H, Korula N, Mirrokni V (2018) Allocation with traffic spikes: Mixing adversarial and stochastic models. *ACM Trans. Econom. Comput.* 6(3–4):1–23.

Even-Dar E, Mannor S, Mansour Y (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Machine Learn. Res.* 7(June):1079–1105.

Gupta A, Koren T, Talwar K (2019) Better algorithms for stochastic bandits with adversarial corruptions. *Proc. Conf. Learn. Theory*.

Han Y, Wang Y, Chen X (2021) Adversarial combinatorial bandits with general non-linear reward functions. *Proc. Internat. Conf. Machine Learn*.

Huber PJ (1964) Robust estimation of a location parameter. *Ann. Math. Statist.* 35(1):73–101.

Huber PJ, Ronchetti EM (2011) *Robust Statistics*, Series in Probability and Statistics (Wiley, New York).

Lykouris T, Mirrokni V, Leme RP (2018) Stochastic bandits robust to adversarial corruptions. *Proc. ACM Sympos. Theory Comput. (STOC)*.

Mahajan S, van Ryzin G (2001) Stocking retail assortments under dynamic consumer substitution. *Oper. Res.* 49:334–351.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

Oh MH, Iyengar G (2019) Multinomial logit contextual bandits. *Reinforcement Learn. Real Life (RL4RealLife) Workshop Internat. Conf. Machine Learn. (ICML)*.

Rusmevichientong P, Shen ZJ, Shmoys D (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Saure D, Zeevi A (2013) Optimal dynamic assortment planning with demand learning. *Manufacturing Service Oper. Management* 15(3):387–404.

van Ryzin G, Mahajan S (1999) On the relationships between inventory costs and variety benefits in retail assortments. *Management Sci.* 45:1496–1509.

Xi Chen is an associate professor in the Department of Technology, Operations, and Statistics at Stern School of Business, New York University. His research interests include statistical machine learning, stochastic optimization, and data-driven operations management.

Akshay Krishnamurthy is a principal research manager at Microsoft Research New York City. His research interests include machine learning and statistics, with a particular focus on interactive learning, contextual bandits, and reinforcement learning.

Yining Wang is an associate professor at the Naveen Jindal School of Management, University of Texas at Dallas. He is generally interested in machine learning and its applications in data-driven operations management, such as dynamic pricing, assortment optimization, capacity management and inventory replenishment.