Assortment Optimization under Unknown MultiNomial Logit Choice Models

Wang Chi Cheung, David Simchi-Levi

Abstract
Motivated by e-commerce, we study the online assortment optimization problem. The seller offers an assortment, i.e., a subset of products, to each arriving customer, who then purchases one or no product from her offered assortment. A customer’s purchase decision is governed by the underlying MultiNomial Logit (MNL) choice model. The seller aims to maximize her revenue and learn the underlying choice model simultaneously. Apart from model uncertainty, the seller often faces resource constraints; when a product is sold, a certain amount of resources is consumed, and the resources cannot be replenished during the sales horizon. The seller is then forced to stop the sales process either when the assortment is exhausted or the revenue earned by the seller, who is uncertain about the model. We assume an uncertain MultiNomial Logit (MNL) model, which is fundamental in the literature.

In this paper, we formulate a model for the online assortment optimization problem, which encompasses choice uncertainty and resource constraints. The seller aims to minimize her regret, i.e., the difference between the revenue earned by an oracle, who knows the underlying choice model, and the revenue earned by the seller, who is uncertain about the model. We assume an uncertain MultiNomial Logit (MNL) model, which is fundamental in the literature. We first propose an efficient policy Online(τ) that incurs a regret \( \tilde{O}(T^{2/3}) \), where \( T \) is the number of customers, and \( \tau \) is the length of the learning phase. Then, we propose a UCB policy with a regret of \( \tilde{O}(\sqrt{T}) \); the UCB policy is not known to be computationally efficient. Both regret bounds are sublinear in the total number of assortments, since we exploit the special structure of MNL choice model to avoid learning all the choice probabilities assortment by assortment.

2 Literature Review and Our Contributions
Offline assortment optimization. The MNL choice model is a fundamental model proposed by [McFadden, 1974], and it has been the building block for many other existing choice models [Ben-Akiva and Lerman, 1985]. Assortment optimization under the MNL choice model has been actively studied. Assuming the knowledge of the underlying MNL choice model, [Talluri and van Ryzin, 2004] propose an efficient algorithm for computing an optimal assortment when there is no resource constraint; [Liu and van Ryzin, 2008] propose an efficient algorithm for computing a mixture of assortments that achieves asymptotic optimality under resource constraints. [Bernstein et al., 2015] offer insights into the optimal assortment planning policy under resource constraints, when the product prices are equal but there are multiple types of customers.

Online assortment optimization. Assuming uncertainty in the MNL choice model, [Rusmevichientong et al., 2010] propose an online policy that incurs an instance-dependent \( O(\log T) \) regret. [Saure and Zeevi, 2013] generalize [Rusmevichientong et al., 2010] by proposing online policies with instance-dependent \( O(\log T) \) regret bounds for a wider class of choice models. Recently, [Agrawal et al., 2016] provide a instance-independent regret \( O(\sqrt{T}) \) under an uncertain MNL choice model. However, these existing works do not incorporate resource constraints into their models, unlike ours. Our approach is based on establishing a confidence bound on the choice probability for every
assortment (cf. Lemmas [S.4] [S.2]), which is novel in the literature, and necessary for learning the choice model under resource constraints.

**Budgeted bandits.** Our online assortment optimization problem can be cast as a budgeted bandit problem, with the arm set being the allowed assortments. For budget bandit problems, [Tran-Thanh et al., 2010] provide an instance-independent regret bound with a resource constraint; [Tran-Thanh et al., 2012] and [Xia et al., 2015a] provide instance-dependent regret bounds for the cases of discrete and continuous resource consumption costs. [Xia et al., 2015b] propose a Thompson Sampling based algorithm. [Badanidiyuru et al., 2013] provide optimal instance-independent regret bounds for the problem with general resource constraints.

A direct application of [Badanidiyuru et al., 2013] or [Agrawal and Devanur, 2014] to our problem yields a regret linear in the number of assortments, which is often larger than the number of customers. Indeed, their policies involve testing each assortment at least once. In contrast, we exploit the special structure of the MNL choice model to achieve a regret bound sublinear in the number of assortments.

**Combinatorial bandits.** Our problem can be cast as a stochastic combinatorial bandit problem with semi-bandit feedback, when we relax the resource constraints, and interpret a product and an assortment as a basic arm and a super arm respectively. [Gai et al., 2012] study the combinatorial bandit problem with linear reward (i.e. a super arm’s reward is the sum of its basic arms’ reward), which is subsequently generalized and refined by [Chen et al., 2013] to the case with non-linear reward. The optimal regret bound is obtained by [Kveton et al., 2014] in the case of linear reward. [Chen et al., 2016] consider the generalized case when the expected reward under a super arm depends on certain random variables associated with its basic arms. [Xia et al., 2016] provide an instance-dependent regret bound to the combinatorial bandit problem with a resource constraint. Recent works [Radlinski et al., 2008], [Kveton et al., 2015a], [Kveton et al., 2015b] consider the problem in the cascading-feedback setting.

Apart from the presence of resource constraints (except [Xia et al., 2016]), our model differs from the existing combinatorial bandit literature, as our reward function is not monotonic in the super arm and the underlying parameters. Indeed, introducing more products in an assortment does not necessarily increase the expected revenue, since the customer’s attention could be diverted to less profitable products. Therefore, novel techniques are needed for achieving a sublinear regret in our setting.

### 3 Problem Definition

We formulate the online assortment optimization problem with an unknown MultiNominal Logit (MNL) choice model. The seller has a set of products $\mathcal{N} = \{1, \ldots, N\}$ for sale, and a set of resources $\mathcal{K} = \{1, \ldots, K\}$ for composing the products. The sale of one product $i$ generates a revenue of $r(i) \in [0, 1]$, but consumes $a(i, k) \in \{0, 1\}$ units of resource $k$, for each $k \in \mathcal{K}$. Product 0 is the “no-purchase” product; $r(0) = 0 = a(0, k)$ for all $k \in \mathcal{K}$.

The seller starts with $C(k) = Tc(k) \in \mathbb{Z}^+$ units of resource $k$ at period 1. For periods $t = 1, \ldots, T$, the following sequence of six events happen. First, a customer arrives in period $t$. Second, the seller offers an assortment $S_t \in \mathcal{S}$ to the customer, where $\mathcal{S}$ is the family of allowed assortments. Third, the seller observes that the product $I_i \in S_t \cup \{0\}$ is purchased. Fourth, the seller earns a revenue of $r(I_i)$. Fifth, the resources are consumed: for all $k \in \mathcal{K}$, $C(k) \leftarrow C(k) - a(I_i, k)$. Sixth, the seller proceeds to period $t + 1$.

A customer’s purchase decision is governed by the MNL choice probability function $\varphi(\cdot)\vert v^* \) [McFadden, 1974], $v^* \in \mathbb{R}_+^N$ is the latent utility parameter unknown to the seller; the seller only knows that $v^*(i) \in [1/R, R]$ for all $i \in \mathcal{N}$. For $i \in \mathcal{S} \subset \mathcal{N}$ and $v \in \mathbb{R}_+^N$, $\varphi(i, S\vert v)$ represents the probability of a customer purchasing $i$ when she is offered assortment $S$, and has utility parameter $v$. The probability is defined as

$$\varphi(i, S\vert v) := \frac{v(i)}{1 + \sum_{i \in S} v(i)}.$$  

The customer purchases nothing with the complementary probability $\varphi(0, S\vert v) = 1/(1 + \sum_{i \in S} v(i)) = 1 - \sum_{i \in S} \varphi(i, S\vert v)$. For $i \in \mathcal{S} \subset \mathcal{N}$, $v \in \mathbb{R}_+^N$, we define $\varphi(i, S\vert v) = 0$. The expected revenue $\sum_{i \in S} r(i)\varphi(i, S\vert v)$ is not monotonic in $S$ or $v$, in contrary to the monotonicity of reward functions in the combinatorial bandit literature.

The family of allowed assortments $\mathcal{S}$ is a subfamily of $2^\mathcal{N}$. One common example is the cardinality constrained family $\mathcal{S} = \{S \subset \mathcal{N} : |S| \leq B\}$. We assume that $\emptyset \in \mathcal{S}$; that is, the seller can reject a customer by offering an empty assortment, for example when the resources are depleted. We denote $B = \max\{|S| : S \in \mathcal{S}\}$ as the maximum assortment size; in most settings, $B$ is much smaller than $N$, the number of products.

**Regret Minimization.** The seller’s objective is to design a non-anticipatory policy that maximizes the total revenue $\sum_{t=1}^T r(I_t)$, subject to the resource constraints. This can be formulated as the minimization of regret, which is

$$\text{REG} = T\text{OPT}(LP(v^*)) - \sum_{t=1}^T r(I_t),$$

subject to the resource constraints: for all $k \in \mathcal{K}$, $\sum_{t=1}^T a(I_t, k) \leq Tc(k) \text{ always}$. Equivalently, we require $C(k) \geq 0$ at every period. The purchased product $I_t$ depends on the offered assortment $S_t$ determined by the policy. We say that a policy is non-anticipatory if the offered assortment $S_t$ depends only on the sales history as well as the seller’s randomness $U_t$ in period $t$, i.e. $S_t \in \sigma(U_t, \{S_0, I_1, \ldots, U_t\}_{t=1}^{t-1})$.

For any $v \in \mathbb{R}_+^N$, the linear program $LP(v)$ is defined as

$$\max \sum_{S \in \mathcal{S}} R(S\vert v) y(S)$$

subject to

$$\sum_{S \in \mathcal{S}} A(S, k\vert v) y(S) \leq c(k) \quad \forall k \in \mathcal{K},$$

$$\sum_{S \in \mathcal{S}} y(S) = 1, \quad y(S) \geq 0 \quad \forall S \in \mathcal{S}. $$


We use the notation $R(S|v) = \sum_{i \in S} r(i) \varphi(i, S|v)$ to denote the expected revenue earned by offering $S$ in a period, and $A(S, k|v) = \sum_{i \in S} a(i, k) \varphi(i, S|v)$ to denote the expected amount of resource $k$ consumed in a period. The optimal value of LP$(v)$ is denoted as $\text{Opt}(\text{LP}(v))$. By interpreting $y$ as a probability distribution over $S$, LP$(v)$ is equivalent to the maximization of the expected revenue in a period, when the resource constraints hold in expectation. LP$(v)$ is always feasible, since $y(\emptyset) = 1$, $y(S) = 0$ for all $S \in \mathcal{S} \setminus \{\emptyset\}$ is a feasible solution. The benchmark $T\text{Opt}(\text{LP}(v^*))$ upper bounds the expected optimum [Badanidiyuru et al., 2013]:

**Theorem 3.1** ([Badanidiyuru et al., 2013]). For any non-anticipatory policy $\pi$ that satisfies the resource constraints with probability $I$, the following inequality holds:

$$T\text{Opt}(\text{LP}(v^*)) \geq \mathbb{E} \left[ \sum_{t=1}^{T} r(I_t^\pi) \right].$$

$I_t^\pi$ denotes the random product purchased by the period $t$ customer under policy $\pi$.

## 4 Online policy ONLINE($\tau$)

We propose the non-anticipatory policy ONLINE($\tau$), where $\tau$ is the length of the learning phase. ONLINE($\tau$) enjoys the following performance guarantee:

**Theorem 4.1.** Suppose $\tau$ satisfies Assumption 4.2. The policy ONLINE($\tau$) satisfies all resource constraints and incurs a regret at most

$$\tau + O \left( TRB \sqrt{\frac{N \log N}{\tau}} + O \left( T \log \frac{K + 1}{\delta} \right) \right)$$

with probability $1 - \delta$. In particular, the choice $\tau = (TRB)^{2/3}N^{1/3}$ minimizes the regret bound up to a constant factor, yielding the bound $O((TRB)^{2/3}N^{1/3})$.

Our regret bound is sublinear in $N, B$, in deep contrast with [Agrawal and Devanur, 2014], which are linear in $|\mathcal{S}| = \Theta(N^B)$. For our theoretical analysis, we assume the following on $\tau$:

**Assumption 4.2.** The learning phase length $\tau$ satisfies: (i) For all $k \in \mathcal{K}$, $\tau \sqrt{\log \frac{4NK}{\delta}} \leq Tc(k)$. (ii) For all $k \in \mathcal{K}$, $C\epsilon(\tau) \leq \frac{1}{2}c(k)$, where

$$\epsilon(\tau) = 4R \sqrt{\frac{N \log 4N}{\tau}}.$$

Assumption 4.2(i) ensures that no resource is depleted during the learning phase, and (ii) ensures that the learning phase is long enough for estimating $v^*$. Assumption 4.2 is only necessary for our analysis; ONLINE($\tau$) can be implemented for any choice of $1 \leq \tau \leq T$. In our simulation results in [9] ONLINE($T^{2/3}$) still converges to optimal, even when the assumption is violated for the choice $\tau = T^{2/3}$. (Theorem 4.1 implies a regret of $O(T^{2/3}R\sqrt{N/T})$ if $\tau = T^{2/3}$ satisfies Assumption 4.2.) We further discuss the assumption in Appendix A.

### Algorithm 1 ONLINE($\tau$)

1. Initialize $C(k) = Tc(k)$ for all $k \in K$.
2. for $i = 1, \ldots, N$ do ▷ Learning Phase
3. for $t = (i - 1)\tau/N + 1$ to $i\tau/N$ do
4. $S_t = \{i\}$, observe outcome $I_t \in \{i, 0\}$.
5. $S_t$ for all $k \in K$, $C(k) \leftarrow C(k) - a(I_t, k)$.
6. end for
7. Compute the MLE $\hat{v}(i) \in \text{argmin}_{v \in [1/R, T]} L_i(v)$.
8. end for
9. Solve LP($\hat{v}$) for an extreme point solution $\hat{y}$.
10. for $t = \tau + 1, \ldots, T$ do ▷ Earning Phase
11. Offer $S_t$ with probability $\hat{y}(S_t)$.
12. Observe outcome $I_t \in S_t \cup \{0\}$.
13. For all $k \in K$, $C(k) \leftarrow C(k) - a(I_t, k)$.
14. if $\exists k \in K$ s.t. $C(k) = 0$ then
15. ABORT; offer $S = \emptyset$ till the end.
16. end if
17. end for

ONLINE($\tau$) is presented in Algorithm 1. Periods 1 to $\tau$ are the learning phase, and periods $\tau + 1$ to $T$ are the earning phase. During the learning phase, the seller offers single item assortments in order to estimate $\{v^*(i)\}_{i \in \mathcal{N}}$. When the learning phase ends, he computes the MLE $\hat{v}(i)$ for each product. $\hat{v}(i)$ is a solution to $\min_{v \in [1/R, T]} L_i(v)$. The negative log likelihood $L_i(v)$ is

$$= -\log \left[ \prod_{s=(i-1)\tau+1}^{i\tau/N} \left( \frac{v}{1 + v} \right)^{1(I_s=i)} \left( \frac{1}{1 + v} \right)^{1(I_s=0)} \right]$$

$$= n(i) \log \left[ 1 + \frac{1}{v} \right] + \left( \frac{\tau}{N} - n(i) \right) \log [1 + v],$$

where $n(i) = \sum_{s=(i-1)\tau+1}^{i\tau/N} 1(I_s=i)$ is the number of product $i$ sold during the learning phase.

After that, we solve LP($\hat{v}$) for an extreme point solution $\hat{y}$, which can be interpreted as a probability distribution over $S$. Finally, in the earning phase, we offer $S \in \mathcal{S}$ with probability $\hat{y}(S)$ each period. At the end of a period, the seller signals ABORT when some resource is depleted, i.e. $C(k) = 0$. Then, the seller offers empty assortments to subsequent customers, until the end of sales horizon. This ensures that the resource constraints are satisfied with probability 1.

**Computational Efficiency of ONLINE($\tau$).** The most computationally onerous step in ONLINE($\tau$) is to solve LP($\hat{v}$), which has $|\mathcal{S}| = \Theta(N^B)$ many variables. Fortunately, by Liu and van Ryzin, 2008, LP($\hat{v}$) can be efficiently solved by the Column Generation algorithm (CG). In each iteration of CG, we solve the reduced problem $\max_{S \in \mathcal{S}} R(S|\hat{v}) = \max_{S \in \mathcal{S}} \sum_{i \in S} \hat{r}(i) \varphi(i, S|\hat{v})$, where $\hat{r}(i)$ is a suitably defined reduced revenue coefficient for $i$. The reduced problem is polynomial time solvable for many choices of $S$, such as $S = \{S : |S| \leq B\}$ [Rusmevichientong et al., 2010]. In our simulations in §8, CG always terminates within 50 iterations for solving LP($\hat{v}$). Finally, the support of $\hat{y}$, which is defined as
supp($\hat{y}$) := \{S ∈ S : \hat{y}(S) > 0\}, has size $\leq K + 1$, since $\hat{y}$ is an extreme point solution to LP($v$). Thus, it is easy to sample $S_t$ in the earning phase.

A $O(\sqrt{T})$ regret policy. A $O(\sqrt{T})$ regret can be achieved by a UCB policy:

**Theorem 4.3.** There exists a UCB policy that satisfies the resource constraints and achieves a regret of $O\left(\sqrt{T R^3 B^{5/2} N \log T N K}\right)$ with probability at least $1 - \delta$.

The design and analysis of such a UCB policy is deferred to Appendices D - G. Different from ONLINE($\tau$), our UCB policy is not known to be empirically efficient.

5 Overview of the Proof for Theorem 4.1

To begin the proof, we consider the period $t_{\text{last}}$ of last sale. Either ABORT is signaled at the end of period $t_{\text{last}}$, or $t_{\text{last}} = T$, $t_{\text{last}}$ is a random variable, depending on the resource consumption in the sales horizon. Denote $\mathcal{B}$ as $\text{BOUND}(\tau)$. We analyze the regret by the following:

\[
P[\text{REG} \leq \text{BOUND}(\tau)] \\
\geq \mathbb{P}\left[ T \text{OPT}(v^*) - \sum_{t=1}^{T} r(I_t) \leq \text{BOUND}(\tau) \right] \\
(\dagger) \geq \mathbb{P}\left[ T \text{OPT}(v^*) - \sum_{t=1}^{T-\rho} r(I_t) \leq \text{BOUND}(\tau), t_{\text{stop}} > T - \rho \right] \\
(\ddagger) \geq \mathbb{P}\left[ T \text{OPT}(v^*) - \sum_{t=\tau+1}^{T-\rho} r(\hat{I}_t) \leq \text{BOUND}(\tau) \right] \\
\bigcap_{k=1}^{K} \{\tau + \sum_{t=\tau+1}^{T-\rho} a(\hat{I}_t, k) \leq T c(k)\} \\
\mathcal{E}_{\text{REG}} \\
(\ddagger) \geq \mathbb{P}\left[ \mathcal{E}_{\text{REG}} \cap \bigcap_{k=1}^{K} \mathcal{E}_k \mid \mathcal{E}_0 \right] \mathbb{P}[\mathcal{E}_0]. \tag{7}
\]

To prove the Theorem, it suffices to show that the probability $\mathcal{E}_0$ is at least $1 - \delta$.

**Parsing the calculation above.** In step $(\dagger)$, we consider the event $t_{\text{last}} \leq \rho$, where $\rho$ is the constant $\rho = \frac{T c(\epsilon(\tau))}{\min_{k \in K} c(k)} + \sqrt{T \log \frac{4(K+1)}{\delta}}$, where $\epsilon(\tau)$ is defined in $[5]$. The definition of $\rho$ is motivated in the subsequent analysis. The inequality $(\dagger)$ is evidently true, since the probability does not increase when we require the additional event $t_{\text{stop}} > T - \rho$ to hold.

To ease the analysis, we decouple the revenue and the constraints at step $(\ddagger)$, by considering the process $\{\hat{S}_t, \hat{I}_t\}_{t=\tau+1}^{T-\rho}$ generated in Procedure 2. The samples $\hat{S}_{\tau+1}, \ldots, \hat{S}_{T-\rho}$ are i.i.d., where $\mathbb{P}[\hat{S}_t = S] = \hat{y}(S)$. The samples $\hat{I}_{\tau+1}, \ldots, \hat{I}_{T-\rho}$ are independent, where $\mathbb{P}[\hat{I}_t = i] = \phi(i, \hat{S}_t|v^*)$.

The process $\{\hat{S}_t, A\}$ is motivated in Appendix B. First, we argue that the MLE $\hat{v}$ is an extreme point solution to LP($\hat{v}^*$). The step $(\ddagger)$ holds, since the probability does not increase when we require the additional event $\mathcal{E}_0$ to hold. The event $\mathcal{E}_0$ is defined as

\[
\left\{ \log \frac{\hat{v}(i)}{v^*(i)} \leq c(\tau) = 4R \frac{N}{\tau} \log \frac{4N}{\delta} \text{ for all } i \right\}. \tag{9}
\]

The event $\mathcal{E}_0$ implies that MLE $\hat{v}$ is an accurate estimator for $v^*$, with the specified confidence radius. Now, we show that the probability $\mathcal{E}_0$ is at least $1 - \delta$. This is the heart of our proof for the regret bound.

**Proving that the probability $\mathcal{E}_0 \geq 1 - \delta$.** This is proved by combining Lemmas 5.1 - 5.5. Their proofs are deferred to Appendix B. First, we argue that the MLE $\hat{v}$ is sufficiently accurate, in the sense that the event $\mathcal{E}_0$ happens with high probability:

**Lemma 5.1.** For any $\tau \geq N$, $\mathbb{P}[\mathcal{E}_0] \geq 1 - \delta/2$.

The proof involves a change of variable $v = e^{\theta}$, and uses the strong convexity of $L_s(\epsilon(\theta))$ in $\theta$. We next bound the probability $\mathbb{P}[\mathcal{E}_{\text{REG}} \cap \bigcap_{k=1}^{K} \mathcal{E}_k \mid \mathcal{E}_0]$ by the following four Lemmas. We translate the accuracy in estimating $v^*$ to the accuracy in estimating the choice probability for every assortment:

**Lemma 5.2.** For all $v, v' \in \mathbb{R}_{>0}^N$, $b \in [0, 1]^N$ and $S \subset N$, the following inequality holds:

\[
\sum_{i \in S} b(i) (\phi(i, S|v) - \phi(i, S|v')) \leq \sum_{i \in S} \left| \log \frac{v(i)}{v'(i)} \right|.
\]
Lemma 5.2 establishes the Lipschitz continuity of $\varphi(t, S|v)$ in $\log v$. Altogether, Lemmas 5.1 and 5.2 demonstrate that the choice probability under every assortment can be learned without testing every assortment. Furthermore, the Lemmas show that $|R(S|v) - R(S|v^*)| = O(1/\sqrt{T})$ for all $S \in \mathcal{S}$, and that $|A(S, k|v) - A(S, k|v^*)| = O(1/\sqrt{T})$ for all $S \in \mathcal{S}, k \in \mathcal{K}$. This leads to the following Lemma:

**Lemma 5.3.** Condition on $\mathcal{E}_i$ (cf. (9)), we have

$$\text{Opt}(LP(\hat{v})) \geq \left[ 1 - \frac{Be(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}} \right] \text{Opt}(LP(v^*)) - Be(\tau).$$

Assumption 4.2(ii) ensures that $\frac{Be(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}} < 1$. Using Lemma 5.3, we first prove the near optimality in revenue:

**Lemma 5.4.** We have $\mathbb{P}[\mathcal{E}_i | \mathcal{E}_i] \geq 1 - \frac{\delta}{2(K+1)}$.

The proof involves a decomposition of the regret in revenue and applications of Chernoff inequality. Finally, we also argue that resource $k$ are not fully consumed before period $T - \rho$.

**Lemma 5.5.** We have $\mathbb{P}[\mathcal{E}_k | \mathcal{E}_i] \geq 1 - \frac{\delta}{2(K+1)}$ for all $k \in \mathcal{K}$.

The proof for Lemma 5.5 is similar to the proof of Lemma 5.4. Altogether, the regret bound in Theorem 4.1 is proved.

### 6 Numerical Experiments

We evaluate the performance of ONLINE($T^{2/3}$) with synthetic data, with varying model parameters. By Theorem 4.1, it incurs a regret $O(T^{2/3}RB\sqrt{N})$. We define a class tuple $\Gamma = (S, N, \mathcal{K}, R)$, and consider random problems model generated based on $(\Gamma_i)_{i=1}^3$ and 8 sales horizon lengths $\{T(q)\}_{q=1}^5$, which are defined below:

$\Gamma_1 = (S_1(6), 10, 5, 3), \quad \Gamma_2 = (S_1(9), 15, 6, 5), \quad \Gamma_3 = (S_1(15), 25, 8, 7),$

$T = [250, 500, 750, 1000, 1500, 2000, 5000, 10000]$.

Here, we denote $S_1(B) = \{S \subset \mathcal{N} : |S| \leq B\}$. The tuples $\Gamma_1, \Gamma_2, \Gamma_3$ are ordered with increasing difficulty; the number of assortments in $\Gamma_1, \Gamma_2, \Gamma_3$ are 210, 5005, and 3.27×10⁶ respectively. In many cases (especially $\Gamma_3$), there are more possible assortments than the number of periods, which makes the existing budgeted bandit policies (cf. §4) infeasible.

For each $(\Gamma_i, T(q))$, we generate 5 random problem models. Then, for each of the problem models, we run ONLINE($T(q)^{2/3}$) 200 times, over the synthetic data generated with the model. After that, for each model, we compute two quantities: (a) the average revenue-to-optimum ratio, which is the earned revenue averaged over the 200 simulation runs divided by $T(q)\text{Opt}(LP(v^*))$, and (b) the average regret, which is $T(q)\text{Opt}(LP(\hat{v}))$ minus the earned revenue averaged over the 200 runs. Finally, for each $(\Gamma_i, T(q))$, we further average the quantities (a, b) over the 5 generated models.

ONLINE($\tau$) is very efficient via the use of CG (cf. §4). In our simulation, CG always terminates in 50 iterations, and each run can be simulated in less than 10 seconds for models from $\Gamma_3$.

Fig. 1a depicts the trend of the average revenue-to-optimum ratio for each $\Gamma_i$ when $T(q)$ varies. The ratio converges to 1 as $T(q)$ increases. In addition, our policy performs well even when $T(q)$ is small. For example, for $\Gamma_3$, where $|S| = 3.27 \times 10^6$, our policy is still able to achieves a ratio of 0.68 when $T(q) = 250$. This demonstrates that ONLINE($T^{2/3}$) performs well even when Assumption 4.2 is violated and the sales horizon is short. After witnessing the convergence, we investigate the regret's growth rate. Fig. 1b depicts the trend of the average regret for each $\Gamma_i$ when $T(q)$ varies, in the log-log scale. The black dashed line represents $f(T) = T^{2/3}$, which has slope $\frac{2}{3}$. Observe that the simulated regret grows at a rate $T^{2/3}$, confirming Theorem 4.1.

To further study the convergence of our policy, we examine the how often ONLINE($T^{2/3}$) correctly identify $\text{supp}(y^*)$ after the learning phase. (Recall the notation $\text{supp}(y) = \{S \in \mathcal{S} : y(S) > 0\}$.) In Table 1, for each class tuple and $T$ we tabulate the fraction of instances, out of 200 runs, where $\text{supp}(\hat{y}) = \text{supp}(y^*)$. This is a stringent criterion, since $\text{supp}(\hat{y})$ could be different from $\text{supp}(y^*)$ because of multiplicity in the optimal solutions for LP($v^*$), and near optimality could still be achieved without $\text{supp}(\hat{y}) = \text{supp}(y^*)$. However, ONLINE($T^{2/3}$) is still able to identify the support in small instances. Additional simulation results show similar trend of convergence and effectiveness in short sales horizon. The details are provided Appendix A.

### 7 Conclusion and Future Directions

The online assortment optimization problem under model uncertainty and resource constraints is studied. We propose online policies, with regret bounds sublinear in the number of periods and assortments. Many interesting research directions remain to be explored. First, it is not known if the regret lower bound by Agrawal et al., 2016 can be attained. Second, the incorporation of contextual information, similar to Chu et al., 2011, Agrawal and Devanur, 2016, is an exciting topic.

### A A Discussion on Assumption 4.2

We remark that the choices of $\tau = T^{2/3}R^{2/3}B^{2/3}N^{1/3}$ and $\tau = T^{2/3}$ satisfy Assumption 4.2 when $T$ is sufficiently large.
Indeed, for the case of $\tau = T^{2/3} R^{2/3} B^{2/3} N^{1/3}$, Assumption 4.2 (i, ii) are equivalent to

$$T \geq \frac{R^2 B^2}{c(k)^3} \log^{3/2} \frac{4NK}{\delta}, \quad T \geq \frac{R^2 B^2 N}{c(k)^3} \log^{3/2} \frac{4N}{\delta}$$

for all $k \in K$. For the case of $\tau = T^{2/3}$, Assumption 4.2 (i, ii) are equivalent to

$$T \geq \frac{1}{c(k)^3} \log^{3/2} \frac{4NK}{\delta}, \quad T \geq \frac{1}{c(k)^3} \log^{3/2} \frac{4N}{\delta}$$

for all $k \in K$. Again for the case of $\tau = T^{2/3}$, our numerical results in [8] shows that ONLINE($\tau$) is effective even when the assumption is violated.

**B. Proofs for the Lemmas in Section 5**

**B.1 Proof of Lemma 5.1**

Recall the definition of $L_i(v)$ in (6). Consider the change of variable $e^{\theta} = v$, and let $L_i(\theta) = L_i(e^{\theta})$. We have

$$L_i(\theta) = n(i) \log \left[ 1 + e^{-\theta} \right] + \frac{\tau}{N} n(i) \log \left[ 1 + e^{\theta} \right].$$

Denote $\tilde{\theta} = \log \tilde{v}(i)$, and $\theta^* = \log \nu^*(i)$. By a Taylor Series Expansion on $f(\gamma) = L_i(\gamma\theta^* + (1 - \gamma)\tilde{\theta})$, we have

$$L_i(\tilde{\theta}) = L_i(\theta^*) + L_i'(\theta^*)(\tilde{\theta} - \theta^*) + 0.5 L_i''(\tilde{\theta})(\tilde{\theta} - \theta^*)^2,$$

where $\tilde{\theta} = \gamma \theta^* + (1 - \gamma)\tilde{\theta}$ for some $\gamma \in (0, 1)$. Since $L_i(\tilde{\theta}) \leq L_i(\theta^*)$, we have

$$0 \geq L_i'(\theta^*)(\tilde{\theta} - \theta^*) + 0.5 L_i''(\tilde{\theta})(\tilde{\theta} - \theta^*)^2. \quad (10)$$

Interestingly, the first derivative term can be bounded as follows:

$$|L_i'(\theta^*)| = \frac{\tau}{N} \frac{e^{\theta^*}}{1 + e^{\theta^*}} - n(i) \leq \sqrt{\frac{\tau}{N} \log \frac{4N}{\delta}} \quad (11)$$

with probability at least $1 - \delta/2N$. (11) is by Chernoff Inequality, since $N(i)$ is a sum of $\tau/N$ i.i.d. 0-1 random variables $\{1(I_s = i)\}_{s=1}^{\tau/N}$, which has expectation $e^{\theta^*}/(1 + e^{\theta^*})$. Next, we bound the second derivative as follows:

$$L_i''(\tilde{\theta}) = \frac{\tau e^{\theta^*}}{N(1 + e^{\theta^*})^2} \geq \frac{R}{N(1 + R)^2}. \quad (12)$$

Combining (10), (11), and substituting $\nu^*(i), \tilde{v}(i), \tilde{\theta}, \theta^*$, we have

$$\frac{\tau R}{2N(1 + R)^2} \log \frac{\tilde{v}(i)^2}{\nu^*(i)} - \sqrt{\frac{\tau}{N} \log \frac{4N}{\delta} \log \frac{\tilde{v}(i)}{\nu^*(i)}} \leq 0.$$

with probability at least $1 - \delta/2N$. Finally, the Lemma is proved by taking a union bound over all products.

**B.2 Proof of Lemma 5.2**

Consider function $f : [0, 1] \to \mathbb{R}$ defined by

$$f(\gamma) = \sum_{i \in S} b(i) (\varphi(i, S) \exp[\theta' + \gamma(\theta - \theta')]),$$

where $\theta' = \{\theta'(i)\}_{i \in N} = \{\log[\nu'(i)]\}_{i \in N}$, and $\theta = \{\theta(i)\}_{i \in N} = \{\log[v(i)]\}_{i \in N}$. Let’s also define the shorthand $\theta_i(i) = \theta'(i) + \gamma(\theta(i) - \theta'(i))$. Note that $\theta_0 = \theta'$ and $\theta_1 = \theta$. By the mean value theorem,

$$\sum_{i \in S} b(i) (\varphi(i, S|\nu) - \varphi(i, S|\nu')) = f(1) - f(0) = f'(\gamma) \quad \text{for some } \gamma \in (0, 1)$$

$$= \sum_{i \in S} b(i)e^{\theta_i(i)} \left( \frac{1}{1 + \sum_{i \in S} e^{\theta_i(i)}} (\theta(i) - \theta'(i)) \right) - \sum_{i \in S} e^{\theta_i(i)} \sum_{\ell \in S} b(\ell)e^{\theta_{\ell}(i)}(\theta(i) - \theta'(i)) \quad (13)$$

$$\leq \sum_{i \in S} |\theta(i) - \theta'(i)| = \sum_{i \in S} \left| \log \frac{\nu(i)}{\nu'(i)} \right|.$$

The inequality (13) holds since the sum of the coefficients of $(\theta(i) - \theta'(i))$ in the two summations lies in $[0, 1]$.

**B.3 Proof of Lemma 5.3**

Consider the following linear program S-LP:

$$\max_{S \subseteq S} [R(S|\tilde{v}) + B\epsilon(\tau)] y(S) \quad (14a)$$

s.t. $\sum_{S \subseteq S} [A(S, k|\tilde{v}) - B\epsilon(\tau)] y(S) \leq c(k) - B\epsilon(\tau) \quad \forall k \in K \quad (14b)$

$$\sum_{S \subseteq S} y(S) = 1, \quad y(S) \geq 0 \quad \forall S \subseteq S, \quad (14c)$$

and let OPT(S-LP) denote its optimal value. We claim the following, conditional on the event $\mathcal{E}_S$:

$$\text{OPT}(LP(\hat{\nu})) + B\epsilon(\tau) \geq \text{OPT}(S-LP) \geq \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in K} \{c(k)\}}\right) \text{OPT}(LP(\nu^*)). \quad (15)$$

**Proving (13):** Rearranging the constraint for resource $k$ yields $\sum_{S \subseteq S} A(S, k|\tilde{v}) y(S) \leq c(k)$, which is the resource $k$ constraint for LP(\tilde{v}). Similarly, the objective of S-LP is equal to the objective of LP(\tilde{v}) plus $B\epsilon(\tau)$. This proves (15).

**Proving (16):** Define the shorthand $\kappa = \frac{B\epsilon(\tau)}{\min_{k \in K} \{c(k)\}}$.

We first claim that the solution

$$\hat{y}(S) = \begin{cases} (1 - \kappa) y^*(S) & \text{if } S \in S \setminus \emptyset \\ \kappa + (1 - \kappa) y^*(\emptyset) & \text{if } S = \emptyset \end{cases}$$

is feasible to S-LP, where $y^*$ is an optimal solution to LP(\nu*). Given the feasibility of $\hat{y}$ to S-LP, we have

$$\text{OPT}(S-LP) \geq \sum_{S \subseteq S} [R(S|\tilde{v}) + B\epsilon(\tau)] \hat{y}(S) \quad (17)$$

$$= \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in K} \{c(k)\}}\right) \sum_{S \subseteq S} R(S|\nu^*) y^*(S)$$
Step (17) is justified as follows. Conditional the event \( \mathcal{E}_\nu \), Lemma [3.5] implies that, for all \( S \in \mathcal{S} \) we have
\[
|R(S|\tilde{y}) - R(S|v^*)| \leq B\epsilon(\tau).
\] (18)
This justifies the step (17).

Finally, we return to checking the feasibility \( \tilde{y} \). First, the constraints in (14) hold; in particular, the equality \( \sum_{S \in \mathcal{S}} \tilde{y}(S) = 1 \) holds by our definition of \( \tilde{y}(\bar{y}) \). Note that the factor \( \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}}\right) \) is non-negative, by Assumption [4.2](ii).

To check the constraints in (14), we have
\[
\sum_{S \in \mathcal{S}} [A(S|\tilde{y}) - B\epsilon(\tau)] \tilde{y}(S) = \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}}\right) \sum_{S \in \mathcal{S}} A(S, k|\tilde{y}) - B\epsilon(\tau) \tilde{y}(S) \leq \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}}\right) \sum_{S \in \mathcal{S}} A(S, k|v^*) \tilde{y}(S)
\] (19)
\[
\leq \left(1 - \frac{B\epsilon(\tau)}{\min_{k \in \mathcal{K}} \{c(k)\}}\right) c(k) \leq c(k) - B\epsilon(\tau),
\] (20)
where (19) is by (18), and (20) is by the feasibility of \( y^* \) to LP(\( v^* \)). Altogether, \( \tilde{y} \) is feasible to S-LP, and this finishes the proof of the Lemma.

B.4 Proof of Lemma [5.4]
Recall the shorthand \( \kappa = B\epsilon(\tau)/\min\{c(k)\} \) used in Appendix B.3. Conditional on \( \mathcal{E}_\nu \), we have:
\[
T\text{Opt}(\text{LP}(v^*)) - \sum_{t=\tau+1}^{T-\rho} \ell(\tilde{I}_t)
\leq T\left(\text{Opt}(\text{LP}(\tilde{y})) + \kappa \text{Opt}(\text{LP}(v^*)) + B\epsilon(\tau)\right) - \sum_{t=\tau+1}^{T-\rho} \ell(\tilde{I}_t)
\leq \rho + \tau + (T - \rho - \tau) (\kappa \text{Opt}(\text{LP}(v^*)) + B\epsilon(\tau))
\] (21)
\[
+ (T - \rho - \tau) \text{Opt}(\text{LP}(\tilde{y})) - \sum_{t=\tau+1}^{T-\rho} \ell(\tilde{I}_t).
\] (22)

The inequality (21) is by Lemma [3.3]. We decompose the term (REGRET) as follows:
\[
\text{(REGRET)} = (T - \rho - \tau) \text{Opt}(\text{LP}(\tilde{y})) - \sum_{t=\tau+1}^{T-\rho} \ell(\tilde{I}_t)
\leq \rho + \tau + (T - \rho - \tau) B\epsilon(\tau)\left(\text{Opt}(\text{LP}(v^*)) + 1\right)
\] (REGRET)
\[
+ T B\epsilon(\tau) + 2 \sqrt{2T \log \frac{4(K + 1)}{\delta}}
\leq \rho + \tau + T B\epsilon(\tau)\left(\frac{1}{\min_{k \in \mathcal{K}} \{c(k)\}} + 2\right)
\] (REGRET)
\[
+ 2 \sqrt{2T \log \frac{4(K + 1)}{\delta}}
\leq T B\epsilon(\tau)\left(\frac{2}{\min_{k \in \mathcal{K}} \{c(k)\}} + 2\right)
\]
\[
+ \left(2 + \frac{1}{\min_{k \in \mathcal{K}} c(k)}\right)\sqrt{2T \log \frac{4(K + 1)}{\delta}} \tag{23}
\]
holds with probability \(1 - \delta/2(K + 1)\). The step (23) is by the definition of \(\rho\) in (8). By the definition of \(\epsilon(\tau)\), the Lemma is proved.

### B.5 Proof of Lemma 5.5

Similar to the proof for Lemma 5.4, we decompose the sum \(\sum_{t=\tau+1}^{T-\rho} a(\tilde{I}_t, k)\) into 4 terms, \((\triangledown_k), (\bigcirc_k), (\bigcirc_k), (\bigotimes_k)\):

\[
\sum_{t=\tau+1}^{T-\rho} a(\tilde{I}_t, k) = \sum_{t=\tau+1}^{T-\rho} a(\tilde{I}_t, k) - A(\tilde{S}_t, k|\hat{v}^*)
\]

\[
\bigcirc_k
\]

\[
+ \sum_{t=\tau+1}^{T-\rho} A(\tilde{S}_t, k|\hat{v}^*) - A(\tilde{S}_t, k|\hat{v})
\]

\[
\bigcirc_k
\]

\[
+ \sum_{t=\tau+1}^{T-\rho} A(S, k|\hat{v}) \hat{y}(S)
\]

\[
\bigotimes_k
\]

To bound \((\bigotimes_k)\): For any fixed sequence of assortments \((\tilde{S}_t)_{t=\tau+1}^{T-\rho}\), the random variables \(a(\tilde{I}_t, k) - A(\tilde{S}_t, k|\hat{v}^*)\) are independent. Each of the random variables \(a(\tilde{I}_t, k) - A(\tilde{S}_t, k|\hat{v}^*)\) has mean zero, and lies in the range \([-1, 1]\). By Chernoff Bound, for any \((\tilde{S}_t)_{t=\tau+1}^{T-\rho}\) the following inequality holds with probability \(1 - \frac{\delta}{4(K + 1)}\):

\[
\sum_{t=\tau+1}^{T-\rho} a(\tilde{I}_t, k) - A(\tilde{S}_t, k|\hat{v}^*) \leq \sqrt{2T \log \frac{4(K + 1)}{\delta}}.
\]

In particular, this is true condition on \(\mathcal{E}_\hat{v}\), hence proving that

\[
\mathbb{P}\left[ (\bigotimes_k) \leq \sqrt{2T \log \frac{4(K + 1)}{\delta}} \mid \mathcal{E}_\hat{v} \right] \geq 1 - \frac{\delta}{4(K + 1)}.
\]

To bound \((\bigcirc_k)\): We bound \((\bigcirc_k)\) in a similar way to the case of \((\bigotimes_k)\). Now, \(a(i, k) \in \{0, 1\}\), and \(|\tilde{S}_t| \leq B\) for all \(t\). Conditional on \(\mathcal{E}_\hat{v}\), we have \(\log \frac{\hat{v}(i)}{v^*(i)} \leq \epsilon(\tau)\). By Lemma 5.2, for all \(i, S\) we have

\[
A(S, k|\hat{v}^*) - A(S, k|\hat{v}) \leq B\epsilon(\tau).
\]

Thus, we have \((\bigcirc_k) \leq TBC(\epsilon(\tau))\).

To bound \((\bigcirc_k)\): Recall that \(\mathbb{P}[\tilde{S}_t = S] = \hat{y}(S)\) (cf. Procedure 4). For any fixed \(\hat{v}\) the random variables

\[
\left\{ A(\tilde{S}_t, k|\hat{v}) - \sum_{S \in \mathcal{S}} A(S, k|\hat{v}) \hat{y}(S) \right\}_{t=\rho+1}^{T-\rho}
\]

are i.i.d., mean 0, and lie in the interval \([-1, 1]\). By Chernoff bound, we have

\[
\mathbb{P}\left[ (\bigcirc_k) \leq \sqrt{2T \log \frac{4(K + 1)}{\delta}} \mid \mathcal{E}_\hat{v} \right] \geq 1 - \frac{\delta}{4(K + 1)}.
\]

To bound \((\bigotimes_k)\): Recall that \((\bigotimes_k) \leq (T - \rho - \tau)c(k), since \(\hat{y}\) is a feasible solution to LP(\(\hat{v}\)).

Altogether, conditional on \(\mathcal{E}_\hat{v}\), the following holds with probability \(1 - \delta/2(K + 1)\):

\[
(\bigotimes_k) + (\bigcirc_k) \leq \sqrt{2T \log \frac{K + 1}{\delta}} + TBc(\epsilon(\tau)) + (T - \rho - \tau)c(k)
\]

\[
\leq Tc(k) - \tau
\]

where (24) is by the definition of \(\rho\) (cf. (8)). Altogether, the Lemma is proved.

### C Additional Simulation Results

We evaluate the performance of ONLINE(\(T^{2/3}\)) with synthetic data, when the family of allowable assortments is a partition matroid. Recall that a class tuple is \((S, N, K, R)\). Define the notation \(\mathcal{S}_2(p, b) = \{S \subset \mathcal{N} : |S \cap \mathcal{N}_j| \leq b\text{ for all } 1 \leq j \leq p\}\), which denotes a partition matroid assortment family. Here, \(\mathcal{N}_1, \ldots, \mathcal{N}_p\) is a partition of \(\mathcal{N}\) into \(p\) equal size subsets, where \(\mathcal{N}_j = \{(N(j - 1)/p) + 1, \ldots, N_j/p\}\). (Thus, we implicitly assume that \(N\) is divisible by \(p\)). By [Davis et al.], the optimization problem \(\max_{S \in \mathcal{S}_2(p, b)} R(S|v)\) is polynomial time solvable, for any \(v, p, b\). Therefore, CG can still be efficiently implemented for ONLINE(\(T^{2/3}\)) (cf. the discussion on the computational efficiency CG in Section 4).

We consider random models generated according to the following class tuples:

\[
\Gamma_4 = (\mathcal{S}_2(2, 3), 10, 5, 3), \quad \Gamma_5 = (\mathcal{S}(3, 3), 15, 6, 5), \quad \Gamma_6 = (\mathcal{S}_2(5, 3), 25, 8, 7).
\]

Similar to Section 6 we evaluate the performance of ONLINE(\(\tau\)) on the problem instances with the following lengths of sales horizon:

\[
T = [250, 500, 750, 1000, 1500, 2000, 5000, 10000].
\]

Our evaluation procedure is completely identical to the procedure in Section 6. Figure 2 and Table 2 have the same interpretation as Figure 1 and Table 1. Evidently, the simulation performance for partition matroid assortment families is similar to the performance for cardinality constrained assortment families.

| Class \(\cap\) \(T\) | 500 | 1000 | 2000 | 5000 | 10000 |
|----------------|-----|------|------|------|-------|
| \(\mathcal{S}_2(2, 2), 6, 5, 3\) | 0.445 | 0.300 | 0.200 | 0.800 | 0.610 |
| \(\mathcal{S}_2(2, 2), 10, 5, 3\) | 0.230 | 0.300 | 0.355 | 0.385 | 0.450 |
| \(\mathcal{S}_2(4, 3), 15, 6, 3\) | 0.140 | 0.115 | 0.190 | 0.230 | 0.325 |
| \(\mathcal{S}_2(4, 3), 20, 9, 5\) | 0.030 | 0.040 | 0.050 | 0.075 | 0.105 |

Table 2: Fraction of instances where supp(\(\hat{y}\)) = supp(\(y^*\)).
D An Online Algorithm with $O(\sqrt{T})$ regret

In this Appendix Section, we propose and analyze a UCB policy that achieves a $O(\sqrt{T})$ regret. The following statement is the full version of Theorem 4.3.

**Theorem D.1.** Assume that $\omega < 1$, where $\omega$ is defined in (27). Algorithm 3 satisfies the resource constraints with probability 1, and achieves a regret of

$$O \left( \sqrt{T R^3 B^{1/2} N \log \frac{TK}{\delta}} \right)$$

with probability at least $1 - \delta$.

The assumption of $\omega < 1$ ensures that the sales horizon is long enough for sufficient learning. We further explain the rationale behind the assumption in the analysis.

We remark that, while the regret bound for the UCB policy (Algorithm 3) has a better dependence on $T$ than ONLINE($\tau$), the former has a poorer dependence on $R, B, N$ than the latter. It is because the UCB policy estimates the underlying utility parameter $\nu^a$ with a stream of assortments $S_1, S_2, \ldots$ (and the corresponding purchase outcomes $I_1, I_2, \ldots$) of arbitrary sizes. Thus, the UCB policy needs to disentangle the dependence between different products in each offered assortment during the estimation. This situation is in contrast to ONLINE($\tau$), which estimates $\nu^a$ by inferring from single item assortments (and the corresponding purchase outcome). Therefore, ONLINE($\tau$) does not need to go through the disentangling process, leading to a better dependence on the parameters $R, B, N$ than our UCB policy. Nevertheless, since ONLINE($\tau$) separates learning from earning, its dependence on $T$ is strictly worse than the UCB policy, which simultaneously learns and earns.

Our UCB policy is stated in Algorithm 3. The signal ABORT ensures that the resource constraints are satisfied with probability 1.

**Algorithm 3 UCB Policy**

1: Initialize $C(k) = Tc(k)$, $k \in K$, and fixed assortments $\{S_i\}_{i=1}^N$ such that $i \in S_i \subseteq S$.
2: for $t = 1, \ldots, N$ do \hspace{1cm} \Comment{Warm Start}
3: \hspace{1cm} Offer $S_t$, and observe $I_t$.
4: \hspace{1cm} For all $k \in K$, $C(k) \leftarrow C(k) - a(I_t, k)$.
5: end for
6: for $t = N + 1, \ldots, T$ do
7: \hspace{1cm} Compute the MLE $\hat{v}_t$ in (25), based on $(S_s, I_s)_{s=1}^{t-1}$.
8: \hspace{1cm} Solve UCB-LP($\hat{v}_t, n_{t-1}, \omega$) for an optimal $\hat{y}_t$.
9: \hspace{1cm} Offer an assortment $S_t \subseteq S$ with probability $\hat{y}_t(S_t)$.
10: \hspace{1cm} Observe the product $I_t$ purchased.
11: \hspace{1cm} For all $k \in K$, $C(k) \leftarrow C(k) - a(I_t, k)$.
12: \hspace{1cm} if $\exists k \in K$ s.t. $C(k) = 0$ then
13: \hspace{1cm} Signal ABORT, break the for-loop and offer $S = \emptyset$ to the remaining customers.
14: \hspace{1cm} end if
15: end for

In the first $N$ periods, we warm-start our estimation on $\nu^a$ by offering assortments containing each of the products. Then, in each of the periods $N + 1, \ldots, T$, we compute the MLE $\hat{v}_t$ for $\nu^a$ using the observed sales history $(S_s, I_s)_{s=1}^{t-1}$:

$$\hat{v}_t = \arg\min_{v \in [1/R, R]^N} \ell_{t-1}(v),$$

where

$$\ell_{t-1}(v) = \sum_{s=1}^{t-1} - \log(\varphi(I_s, S_s|v)).$$

After that, we solve the following UCB-LP($\hat{v}_t, n_{t-1}, \omega$)

$$\max \sum_{S \in S} \left( R(S \mid v_t) + \sum_{i \in S} \varepsilon(n_{t-1}(i)) \right) y(S)$$

s.t. $\sum_{S \in S} \left( A(S \mid v_t) - \sum_{i \in S} \varepsilon(n_{t-1}(i)) \right) y(S)$

$$\leq (1 - \omega)c(k) \quad \forall k \in K$$

$$\sum_{S \in S} y(S) = 1, \quad y(S) \geq 0 \quad \forall S \in S$$

for an optimal solution $\hat{y}_t$. The parameters in UCB-LP($v_t, n_{t-1}, \omega$) are defined as follows.

$$\varepsilon(n) = (\sqrt{N + 1})\Psi \sqrt{n},$$

$$n_{t-1}(i) = \sum_{s=1}^{t-1} 1(i \in S_t),$$

$$\Psi = R(1 + B)^2 \sqrt{6 \log \frac{2NT(K + 1)}{\delta}},$$

$$\omega = \frac{11\Psi N}{\min_{k \in K} c(k)} \sqrt{\frac{B}{T}} \log \frac{4(K + 1)}{\delta}.$$
Procedure 4 Generation of \( \{ \hat{S}_t, \hat{I}_t \}_{t=1}^{T} \)

1: for \( t = 1, \ldots, N \) do
2: Define \( \hat{S}_t = S_t \), where \( S_t \) are the fixed assortments defined in Line \( 4 \) in Alg \( 3 \)
3: Sample \( \hat{I}_t \sim \hat{S}_t \).
4: end for
5: for \( t = N + 1, \ldots, T \) do
6: Compute the MLE \( \hat{v}_t \) in (29), based on \( \{(\hat{S}_t, \hat{I}_t)\}_{s=1}^{T} \).
7: Solve UCB-LP(\( \hat{v}_t, \hat{n}_{t-1}, \omega \)) for an optimal \( \hat{y}_t \).
8: Select the sample assortment \( \hat{S}_t \) with prob. \( \hat{y}_t(\hat{S}_t) \).
9: Sample \( \hat{I}_t \sim \hat{S}_t \).
10: end for

By the assumption of \( \omega < 1 \) in Theorem D.1, the right hand sides of the constraints (27a) are positive. The linear program UCB-LP(\( v_t, n_{t-1}, \omega \)) is always feasible, since \( y(\emptyset) = 1 \), \( y(S) = 0 \) for all \( S \in S \setminus \{ \emptyset \} \) is always a feasible solution. Different from LP(\( \hat{v} \)) (which is used in ONLINE(\( \tau \))), it is not known if UCB-LP(\( v_t, n_{t-1}, \omega \)) can be efficiently solved (at least empirically) by the Column Generation algorithm or any other algorithm or heuristic.

The incorporation of confidence bounds into UCB-LP(\( v_t, n_{t-1}, \omega \)) is inspired by [Agrawal and Devanur, 2014] as well as the primal-dual algorithm in [Badanidiyuru et al., 2013]. However, our design of the confidence bounds and the analysis are substantially different. As remarked in the design of ONLINE(\( \tau \)). We cannot afford to learn all the choice probabilities \( \{\varphi(i, S|v^*)\}_{i \in N, S \in S} \) individually, which would be the case if we just directly apply [Agrawal and Devanur, 2014] [Badanidiyuru et al., 2013]. Instead, we need to first provide a confidence bound on \( v^* \), and then translate it to corresponding confidence bounds for the choice probabilities. The curse of dimensionality in learning is thus avoided. Different from ONLINE(\( \tau \)), the confidence bounds for the UCB policy is adaptively defined every period.

In the following, we outline the proof of Theorem D.1 in Appendix E and then prove the auxiliary Lemmas and Theorem in Appendices F - H.

E  Proving the \( \tilde{O}(\sqrt{T}) \) Regret

First, we note that one particular challenge in analyzing the UCB policy is that it is ABORTs at the random period \( \tau \) when \( C(k) = 0 \). This makes the analysis of total revenue earned difficult. This is similar to the difficulty in analyzing ONLINE(\( \tau \)).

Thus, to facilitate the analysis, we consider the following sales process \( \{(\hat{S}_t, \hat{I}_t)\}_{t=1}^{T} \) generated by Procedure 4. In Procedure 4, the notation \( \hat{I}_t \sim \hat{S}_t \) denotes sampling a product \( \hat{I}_t \) from \( \hat{S}_t \cup \{ \emptyset \} \) with the underlying choice probability \( \varphi(\hat{I}_t, \hat{S}_t|v^*) \). We define \( n_{t-1}(i) = \sum_{t=1}^{T-1} \min \{1(i \in S_t) \} \), similar to the definition of \( n_{t-1}(i) \) in (29). We emphasize that \( \{(\hat{S}_t, \hat{I}_t)\}_{t=1}^{T} \) is only used for the analysis; the online algorithm does not need to know how to generate such a process. This is similar to the use of Procedure 2 for analyzing ONLINE(\( \tau \)).

Note that \( \{(\hat{S}_t, \hat{I}_t)\}_{t=1}^{T} \) is closely related to the sale process \( \{(S_t, I_t)\}_{t=1}^{T} \) generated by Algorithm 3. Let \( t_{\text{stop}} \) be the period when Algorithm 3 signals ABORT; define \( t_{\text{stop}} = T \) if no ABORT is signaled. When Algorithm 3 does not signal ABORT, the processes \( \{(S_t, I_t, \hat{v}_t, \hat{n}_t)\}_{t=1}^{T} \) and \( \{(\hat{S}_t, \hat{I}_t, \hat{v}_t, \hat{n}_t)\}_{t=1}^{T} \) are identically distributed. However, if an ABORT is signaled at period \( t_{\text{stop}} \), then \( \{(\hat{S}_t, \hat{I}_t)\}_{t=t_{\text{stop}}+1}^{T} \). Moreover, our UCB policy satisfies the resource constraints with probability 1, i.e. \( \sum_{t=1}^{T} a(I_t, k) \leq Tc(k) \leq C(k) \) with certainty; but \( \sum_{t=1}^{T} a(\hat{I}_t, k) > Tc(k) \) violate the constraints with positive (despite being exponentially small) probability.

Now, we have for any target regret bound \( \text{BOUND} \) the following inequality:

\[
\mathbb{P} \left[ \text{Regret} \leq \text{BOUND} \right] = \mathbb{P} \left[ \sum_{t=1}^{T} r(I_t) \leq \text{BOUND} \right] \geq \mathbb{P} \left[ \sum_{t=1}^{T} r(\hat{I}_t) \leq \text{BOUND} \right] \cap \left\{ \sum_{k=1}^{T} a(\hat{I}_t, k) \leq Tc(k) \text{ for all } k \right\}.
\]

To prove Theorem D.1, it suffices to prove the following two Lemmas:

**Lemma E.1.** We have

\[
\mathbb{P} \left[ \sum_{t=1}^{T} r(I_t) \leq \text{BOUND} \right] \geq 1 - \frac{\delta}{K + 1}.
\]

**Lemma E.2.** We have

\[
\mathbb{P} \left[ \sum_{k=1}^{T} a(\hat{I}_t, k) \leq Tc(k) \text{ for all } k \in K \right] \geq 1 - \frac{K \delta}{K + 1}.
\]

The remaining exposition focuses on proving Lemmas E.1 and E.2. To accomplish these tasks, we first prove the following instrumental Theorem, sheds light on the choice of parameter in UCB-LP(\( v_t, n_{t-1}, \omega \)).

**Theorem E.3.** Let \( E_t \) denote the event that the inequality

\[
B(S|\hat{v}_t) - \sum_{i \in S} \varepsilon(\hat{n}_{t-1}(i)) \leq B(S|v^*)
\]

\[
\leq B(S|\hat{v}_t) + \sum_{i \in S} \varepsilon(\hat{n}_{t-1}(i))
\]

holds for all \( S \in S \), \( b \in [0,1]^N \). (We use the notation \( B(S|v) = \sum_{i \in S} b(i) \varphi(i,S|v) \)). Then \( E_t \) holds with probability at least \( 1 - \delta/(2(K+1)T) \).
The proof of Theorem E.3 is first outlined in Appendix E.1 and the main Theorem in Appendix E.1 is proved in Appendix E.2.

Finally, we prove Lemmas E.1, E.2 by using Theorem E.3 in Appendices F, G.

E.1 Proving Theorem E.3

We prove Theorem E.3 by establishing a confidence bound for estimating $v^*$, using correlated samples $\{\tilde{S}_i, \tilde{I}_i\}_{i=1}^{t-1}$ generated by Algorithm 3. Note that $\tilde{S}_i \in \sigma(\{\tilde{S}_j, \tilde{I}_j\}_{j=1}^{t-1} \cup \tilde{U}_t)$, where $\tilde{U}_t$ is the randomness used to generate $\tilde{S}_i$ in Line 8.

**Theorem E.3.** Consider the sales process $\{\tilde{S}_i, \tilde{I}_i\}_{i=1}^{t-1}$ generated by Algorithm 3 where $t \geq N + 1$. The following inequality

$$\sum_{i=1}^{N} \left( \frac{\sqrt{n_{t-1}(i)} \log \frac{\tilde{v}(i)}{v^*(i)} - \Psi}{\sqrt{n_{t-1}(i)} \log \frac{\tilde{v}(i)}{v^*(i)}} \right)^2 \leq N \Psi^2$$

holds with probability at least $1 - \delta / 2T(K + 1)$.

The proof of Theorem E.3 is postponed to Appendix E. The proof is similar to the proof of Lemma E.1, but the analysis in the proof of Theorem E.3 is significantly more involved, since we need to disentangle the dependence across different products for the estimation of $v^*$.

Similar to the proof of Lemma E.1, we consider the following change in variables $v(i) = e^{\theta(i)}$ and the function $L_t(\theta) = L_t((\theta(1), \ldots, \theta(N))) = L_t(e^{\theta(1)}, \ldots, e^{\theta(N)})$. The constant $\Psi$ is a artifact of the strong convexity of $L_t$ in $\theta$. A crucial part of the proof involves demonstrating the concentration property of $\nabla L_t(\theta^*)$, the gradient of $L_t$ at $\theta^* = (\theta^*(i))_{i \in \mathcal{N}} = (\log v^*(i))_{i \in \mathcal{N}}$. However, the classical Azuma-Hoeffding or Chernoff inequality is not directly applicable, since the frequency $n_{t-1}(i)$ is a random variable that correlates with $\{(\tilde{S}_i, \tilde{I}_i)\}_{i=1}^{t-1}$. This is in contrast to the analysis in the proof of Lemma E.1 where the number of observations on product $i$ is fixed to be $\tau/N$. Thus, we employ the following concentration inequality, which is commonly used in the multi-armed bandit literature:

**Lemma E.5.** ([Abbasi-yadkori et al., 2011], [Bubeck et al., 2011]).

Let $\{\mathcal{F}_t\}_{t=1}^\tau$ be a filtration. Let $\rho(t) \in (0, 1)$ be a binary $\mathcal{F}_{t-1}$-measurable random variable, and let $\eta(t)$ be a $\mathcal{F}_t$-measurable random variable that is conditionally centered and $\mathcal{F}_{t-1}$-conditionally $L$-subGaussian, i.e., $\mathbb{E}[\eta(t) | \mathcal{F}_{t-1}] = 0$ a.s. and $\mathbb{E}[\eta^2(t) | \mathcal{F}_{t-1}] \leq c(L\rho(t)^2/2$ for all $\lambda \in \mathbb{R}$.

Then the confidence bound

$$\sum_{t=1}^\tau \rho(t) \eta(t) \leq \mathcal{L} \sqrt{\left(1 + \sum_{t=1}^\tau \rho(t)^2 \right) \left(2 + 2 \log \frac{\tau}{\mathcal{L}}\right)}$$

holds with probability at least $1 - \delta$.

The Lemma follows from either the application of Doob’s Optional Sampling Theorem with Azuma-Hoeffding inequality (for example see the proof of Lemma 15 in [Bubeck et al., 2011]), or from the theory of self-normalizing processes (for example, see Lemma 6 in [Abbasi-yadkori et al., 2011]).

In particular, Theorem E.3 implies that the confidence bound

$$\left| \log \frac{\tilde{v}(i)}{v^*(i)} \right| \leq \varepsilon(n_{t-1}(i))$$

holds for all product $i \in \mathcal{N}$ with probability at least $1 - \delta / 2T(K + 1)$.

Finally, combining (33) with Lemma E.2 Theorem E.3 is proved.

F Proof of Theorem E.4

Recall that $L_{t-1}(v)$ in the Theorem is the negative log-likelihood under the samples $\{\tilde{S}_i, \tilde{I}_i\}_{i=1}^{t-1}$ generated by Procedure 4 (cf. (26)). Consider the the following change of variables and transformation on the likelihood function:

For all $i \in \mathcal{N}$, $\theta(i) = \log v(i)$,

$$L_t(\theta) = L_t((\theta(1), \ldots, \theta(N))) = L_t(e^{\theta(1)}, \ldots, e^{\theta(N)}).$$

Also, we denote $\tilde{\theta}_t = (\log \tilde{v}(i))_{i=1}^{N}$, and $\theta^* = (\log v^*(i))_{i=1}^{N}$.

By Taylor approximation, we know that there exists $\gamma \in [0, 1]$ such that

$$L_{t-1}(\hat{\theta}_t) = L_{t-1}(\theta^*) + \nabla L_{t-1}(\theta^*)^T (\hat{\theta}_t - \theta^*) + \frac{1}{2} (\hat{\theta}_t - \theta^*)^T H_{t-1}(\theta^* + \gamma(\hat{\theta}_t - \theta^*))(\hat{\theta}_t - \theta^*),$$

where

$$\nabla L_{t-1}(\theta^*) = \left( \frac{\partial L_{t-1}(\theta)}{\partial \theta(i)} \right)_{i=1}^{N} \bigg|_{\theta = \theta^*}$$

is the gradient at $\theta^*$, and

$$H_{t-1}(\theta^* + \gamma(\hat{\theta}_t - \theta^*)) = \left( \frac{\partial^2 L_{t-1}(\theta)}{\partial \theta(i) \partial \theta(j)} \right)_{1 \leq i, j \leq N} \bigg|_{\theta = \theta^* + \gamma(\hat{\theta}_t - \theta^*)}$$

is the Hessian matrix.

Now, we know that $L_{t-1}(\theta^*) \geq L_{t-1}(\hat{\theta}_t)$, since $\tilde{v}_t$ minimizes $L_t$. This yields:

$$\nabla L_{t-1}(\theta^*)^T (\hat{\theta}_t - \theta^*) + \frac{1}{2} (\hat{\theta}_t - \theta^*)^T H_{t-1}(\theta^* + s(\hat{\theta}_t - \theta^*))(\hat{\theta}_t - \theta^*) \leq 0.$$

(37)

Now, we claim the following two inequalities:

1. With probability at least $1 - \delta / (2T(K + 1))$, we have

$$\left| \log \frac{\tilde{v}(i)}{v^*(i)} \right| \leq \varepsilon(n_{t-1}(i))$$

holds for all $i \in \mathcal{N}$.

2. We have

$$H_{t-1}(\theta) \geq \frac{1}{R(1 + BR)^2}$$

for all $i \in \mathcal{N}$.
for all $\theta \in [-\log R, \log R]^N$. The notation $A \succeq B$ means that $A - B$ is positive semi-definite. If (38), (39) hold, then we have the following from (37) \[
abla_{\theta} L_{t-1}(\theta^*) - \sum_{i=1}^N \sqrt{n_{t-1}(i)} (\hat{\theta}_i(i) - \theta^*(i)) \leq 0.
\] for all $\theta \in [-\log R, \log R]^N$. The partial derivative has the following expression: \[
abla_{\theta} L_{t-1}(\theta^*) = -\sum_{s \in \{1, \ldots, t-1\}: \tilde{S}_s \ni i} e^{\theta(i)} e^{\theta(j)} \left(1 + \sum_{t \in \tilde{S}_i} e^{\theta(t)}\right)^2.
\] (40) for all $i \in \mathcal{N}$ with probability at least $1 - \delta/(2T(K + 1))$, and the inequality (40) is by the assumption that $\tilde{n}_{t-1}(i) \geq 1$ for all $i \in \mathcal{N}$.

**Proving (39)**. First we express the second derivatives for $L_{t-1}$ for any $\theta \in \mathbb{R}^N$. For $i \neq j$, \[
abla_{\theta} L_{t-1}(\theta^*) = -\sum_{s \in \{1, \ldots, t-1\}: \tilde{S}_s \ni i} e^{\theta(i)} e^{\theta(j)} \left(1 + \sum_{t \in \tilde{S}_i} e^{\theta(t)}\right)^2.
\] (41)
\[
abla_{\theta}^2 L_{t-1}(\theta^*) = \sum_{s \in \{1, \ldots, t-1\}: \tilde{S}_s \ni i} e^{\theta(i)} + \sum_{t \in \tilde{S}_i} e^{\theta(t)} \left(1 + \sum_{t \in \tilde{S}_i} e^{\theta(t)}\right)^2.
\] (42)

Now, we focus on the Hessian matrix $h_s(\theta)$ for the $s^{th}$ period sample $(\tilde{S}_s, \tilde{I}_s)$. We have $H_{t-1}(\theta) = \sum_{i=1}^{t-1} h_s(\theta)$. By (41), (42), the Hessian matrix $h_s(\theta)$ can be expressed as follows:

$$h_s(\theta) = \frac{1}{(1 + \sum_{i \in \tilde{S}_s} e^{\theta(i)})^2} \left( e^{\theta(1)} (1 \in \tilde{S}_s) 0 0 0 0 0 0 0 \\ e^{\theta(2)} (2 \in \tilde{S}_s) 0 0 0 0 0 0 0 \\ \vdots 0 0 0 0 0 0 0 0 \\ e^{\theta(N)} (N \in \tilde{S}_s) \right) + \sum_{1 \leq i < j \leq N: \tilde{S}_i \ni \tilde{S}_j} e^{\theta(i)+\theta(j)} u_{i,j}^T u_{i,j},$$

(43)

where the vector $u_{i,j} = e_i - e_j$, and $e_i$ is the $i^{th}$ standard basis vector. Now, each term in the second summation is positive semi-definite. Applying the bound $\nu = e^{\theta(i)} \in [-R, R]$ for all $i \in \mathcal{N}$, and the model assumption that $|S| \leq B$ for all $S \in \mathcal{S}$, we have \[
abla_{\theta} L_{t-1}(\theta^*) \geq \frac{1}{(1 + \sum_{i \in \tilde{S}_s} e^{\theta(i)})^2} \left( e^{\theta(1)} (1 \in \tilde{S}_s) 0 0 0 0 0 0 0 \\ e^{\theta(2)} (2 \in \tilde{S}_s) 0 0 0 0 0 0 0 \\ \vdots 0 0 0 0 0 0 0 0 \\ e^{\theta(N)} (N \in \tilde{S}_s) \right),
\] (43)

and summing the inequality (43) over $1 \leq s \leq t-1$ yields (39). This concludes the proof of Theorem 4.

**G Proofs of Lemmas** E.1, E.2

**G.1 Proof of Lemma E.1**

To upper bound the regret, we first have the following:

$$(1 - \omega)(T - N) \text{OPT}(LP(\nu^*)) - \sum_{t=N+1}^{T} r(\tilde{I}_t)$$

(44)
for all uncertainty algorithm.

The inequality (47) holds with probability at least $1 - \delta/(2(K + 1))$, by Theorem 5.2 and a union bound over the periods. All inequalities apart from (47) hold with probability 1. The inequality (48) is based on the following observation.

To bound $(\diamondsuit_0)$: By the definition of $\varepsilon$ in (28), we have

$$\sum_{i \in S} r(i) \varphi(i, S|\tilde{v}_t) + \varepsilon(\tilde{n}_{t-1}(i)) \leq 2B\sqrt{N}\Psi$$

for all $i \in  S$ and all $t$. Observe that the $i$th summand $\text{rev}_i = \sum_{s \in S} \sum_{\tilde{v}_t} r(i) \varphi(i, S|\tilde{v}_t) + \varepsilon(\tilde{n}_{t-1}(i))$ appears in (47) at each of the time indexes $\tilde{y}_t$, and it is clear that $\tilde{n}_{t+1}(i) = \tilde{n}_{t+1}(i) + 1$. The inequality (49) is by Jensen’s Inequality. Finally, the inequality (50) is by the fact that at most $B$ products can be included in each of the $T$ assortments.

To bound $(\diamondsuit_0)$: By the definition of $\varepsilon$ in (28), we have

$$\sum_{i \in S} r(i) \varphi(i, S|\tilde{v}_t) + \varepsilon(\tilde{n}_{t-1}(i)) \leq 2B\sqrt{N}\Psi$$

with probability at least $1 - \delta/4(K + 1)$.

The proof of Claim G.1 is given in Appendix H. The claim shows that our UCB policy can indeed be seen as an optimism-in-face-of-uncertainty algorithm.

To bound $(\diamondsuit_0)$: By our model assumption, $r(i) \leq 0, 1$ for all $i \in N$. Observe that the $i$th summand $\text{rev}_i = \sum_{s \in S} r(i) \varphi(i, S|\tilde{v}_t) - r(\tilde{I}_t) \in [-1, 1]$ is a martingale difference with respect to the filtration $\mathcal{F}_t = \sigma\{(\tilde{S}_s, \tilde{I}_s)\}_{s=1}^{t} \cup \{\tilde{S}_{t+1}\}$, in the sense that $\text{rev}_i$ is $\mathcal{F}_i$ measurable, and $\mathbb{E}[\text{rev}_i | \mathcal{F}_{t-1}] = 0$. By applying Azuma-Hoeffding inequality, we have

$$(\diamondsuit_0) \leq 2\sqrt{N}\Psi \sqrt{\frac{4(K + 1)}{\delta}}$$

with probability at least $1 - \delta/4(K + 1)$.

To bound $(\diamondsuit_0)$: We have the following bound:

$$(\diamondsuit_0) \leq \sum_{t \in N+1} \sum_{i \in S} \varepsilon(\tilde{n}_{t-1}(i)) \text{ w. p.} \geq 1 - \frac{\delta}{2(K + 1)}$$

(47)

$$(\diamondsuit_0) \leq 2\Psi(\sqrt{N} + 1) \sum_{i=1}^{N} \frac{1}{\sqrt{n}}$$

(48)
\[ (1 - \omega) \sum_{S \in S} \left( \sum_{i \in S} a(i, k) \varphi(i, S \mid \tilde{v}_t) - \varepsilon(\tilde{n}_{t-1}(i)) \right) y^*(S) \]
\[ \leq (1 - \omega) \sum_{S \in S} \left( \sum_{i \in S} a(i, k) \varphi(i, S \mid \tilde{v}_t) - \varepsilon(\tilde{n}_{t-1}(i)) \right) y^*(S) \]
\[ \leq (1 - \omega) c(k), \]

where inequality (52) is by the fact that \( a(0, k) = 0 \) for all \( k \in \mathcal{K} \), inequality (53) is by the definition of event \( E_t \). (Recall \( E_t \) from Theorem 2.3.)

Since \( \tilde{y}_t \) is optimal for UCB-LP(\( \tilde{v}_t, \tilde{n}_{t-1}, \omega \)), we have
\[ \text{Opt(UCB-LP(\( \tilde{v}_t, \tilde{n}_{t-1}, \omega \)))} \]
\[ = \sum_{S \in S} \left( \sum_{i \in S} r(i) \varphi(i, S \mid \tilde{v}_t) + \varepsilon(\tilde{n}_{t-1}(i)) \right) \tilde{y}_t(S) \]
\[ \geq \sum_{S \in S} \left( \sum_{i \in S} r(i) \varphi(i, S \mid \tilde{v}_t) + \varepsilon(\tilde{n}_{t-1}(i)) \right) \tilde{y}(S) \]
\[ \geq \sum_{S \in S} \left( \sum_{i \in S} r(i) \varphi(i, S \mid v^*) \right) \tilde{y}(S) \]
\[ = (1 - \omega) \text{Opt(LP}(v^*)). \]

Inequality (54) is by the feasibility of \( \tilde{y} \) to UCB-LP(\( \tilde{v}_t, \tilde{n}_{t-1}, \omega \)), and inequality (55) is by the definition of \( E_t \). This proves the Theorem.

### H.1 Proof of Lemma 2.2

For each \( k \in \mathcal{K} \), we have the following:
\[ \sum_{N+1}^{T} a(\tilde{I}_t, k) \]
\[ = \sum_{t=N+1}^{T} a(\tilde{I}_t, k) - \sum_{t=N+1}^{T} \sum_{i \in S_t} a(i, k) \varphi(i, \tilde{S}_t | v^*) \]
\[ (\bigtriangleup_k) \]
\[ + \sum_{t=N+1}^{T} \sum_{i \in S_t} a(i, k) \varphi(i, \tilde{S}_t | v^*) \]
\[ (\bigcirc_k) \]
\[ - \sum_{t=N+1}^{T} \sum_{i \in S_t} a(i, k) \varphi(i, \tilde{S}_t | \tilde{v}_t) - \varepsilon(\tilde{n}_{t-1}(i)) \]
\[ (\bigcirc_k) \]
\[ + \sum_{t=N+1}^{T} \sum_{i \in S_t} a(i, k) \varphi(i, \tilde{S}_t | \tilde{v}_t) - \varepsilon(\tilde{n}_{t-1}(i)) \]
\[ (\bigcirc_k) \]
\[ = \frac{2N}{\delta} \sum_{t=N+1}^{T} \sum_{i \in S_t} a(i, k) \varphi(i, S_t | \tilde{v}_t) - \varepsilon(\tilde{n}_{t-1}(i)) \tilde{y}_t(S) \]
\[ (\bigtriangledown_k) \]

To bound (\( \bigtriangledown_k \)): By our model assumption, \( a(i, k) \in \{0, 1\} \) for all \( i \in \mathcal{N}, k \in \mathcal{K} \). Observe that the \( i \)-th summand \( \varepsilon(\tilde{n}_{t-1}(i)) \) is a martingale difference sequence with respect to the filtration \( \mathcal{F}_t = \sigma(\{\tilde{S}_t, I_s\}_{s=1}^{t} \cup \{\tilde{S}_{t+1}\}) \), in the sense that \( \varepsilon(\tilde{n}_{t-1}(i)) \) is \( \mathcal{F}_t \)-measurable, and \( \mathbb{E}[\varepsilon(\tilde{n}_{t-1}(i)) | \mathcal{F}_{t-1}] = 0 \). By applying Azuma-Hoeffding inequality, we have
\[ \mathbb{E}[\varepsilon(\tilde{n}_{t-1}(i)) | \mathcal{F}_{t-1}] = 0 \]
\[ \leq 2T \log \frac{4(K+1)}{\delta} \]

with probability at least \( 1 - \delta/\sqrt{4(K+1)} \).

To bound (\( \bigtriangleup_k \)): We have
\[ \mathbb{E}[\tilde{y}_t(S) | \mathcal{F}_{t-1}] = 0 \]

w H.1 Proof of Lemma 2.2
[Agrawal and Devanur, 2014] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In ACM Conference on Economics and Computation, 2014.

[Agrawal and Devanur, 2016] Shipra Agrawal and Nikhil R Devanur. Linear contextual bandits with knapsacks. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 3450–3458, 2016.

[Agrawal et al., 2016] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. A near-optimal exploration-exploitation approach for assortment selection. In Proceedings of the 2016 ACM Conference on Economics and Computation, EC ’16, Maastricht, The Netherlands, July 24-28, 2016, pages 599–600, 2016.

[Badanidiyuru et al., 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In 54th Annual IEEE Symposium on Foundations of Computer Science, 2013, 26-29 October, 2013, Berkeley, CA, USA, pages 207–216, 2013.

[Ben-Akiva and Lerman, 1985] Moshe Ben-Akiva and Steven Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press Series in Transportation Studies. MIT Press, 1985.

[Bernstein et al., 2015] Fernando Bernstein, A. Gürhan Kök, and Lei Xie. Dynamic assortment customization with limited inventories. Manufacturing & Service Operations Management, 17(4):538–553, 2015.

[Bubeck et al., 2011] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-arm bandits. J. Mach. Learn. Res., 12:1655–1695, July 2011.

[Chen et al., 2013] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework, results and applications. ICML, 2013.

[Chen et al., 2016] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In Advances in Neural Information Processing Systems 29, pages 1659–1667, 2016.

[Chu et al., 2011] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoffs functions. volume 15, pages 208–214, 2011.

[Davis et al., 2012] James Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment planning under the multinomial logit model with totally unimodular constraint structures. Manuscript.

[Gai et al., 2012] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bands with linear rewards and individual observations. IEEE/ACM Transactions on Networking (TON), 20(5):1466–1478, 2012.

[Kveton et al., 2014] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In AISTATS, 2014.

[Kveton et al., 2015a] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In ICML-15, 2015.

[Kveton et al., 2015b] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Combinatorial cascading bandits. In Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS’15, pages 1450–1458, 2015.

[Liu and van Ryzin, 2008] Qian Liu and Garrett van Ryzin. On the choice-based linear programing model for network revenue management. Manufacturing & Service Operations Management, 10(2), April 2008.

[McFadden, 1974] Daniel McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics, pages 105–142. Academic Press, 1974.

[Radlinski et al., 2008] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. ICML, 2008.

[Rusmevichientong et al., 2010] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. Operations Research, 58(6):1666–1680, 2010.

[Saure and Zeevi, 2013] Denis Saure and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. Manufacturing & Service Operations Management, 15(3):387–404, 2013.

[Talluri and van Ryzin, 2004] Kalyan T. Talluri and Garrett J. van Ryzin. Revenue management under a general discrete choice model of consumer behavior. Management Science, 50(1):15–33, 2004.

[Tran-Thanh et al., 2010] Long Tran-Thanh, Archie C. Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. Epsilon-first policies for budget-limited multi-armed bandits. In AAAI, 2010.

[Tran-Thanh et al., 2012] Long Tran-Thanh, Archie C. Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada., 2012.

[Xia et al., 2015a] Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In Proceedings of the 7th Asian Conference on Machine Learning, ACML 2015, Hong Kong, November 20-22, 2015., pages 317–332, 2015.

[Xia et al., 2015b] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pages 3960–3966, 2015.

[Xia et al., 2016] Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits
with multiple plays. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pages 2210–2216, 2016.