Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

DYNAMIC ASSORTMENT OPTIMIZATION IN LIVE-STREAMING SALES

Zishi Zhang[1], Haidong Li[2], Ying Liu[2], and Yijie Peng[1]

[1]Wuhan Inst. for Artificial Intelligence & Guanghua Sch. of Mgmt., Peking University, Beijing, CHINA
[2]Sch. of Economics and Mngt., University of Chinese Academy of Sciences, Beijing, CHINA

ABSTRACT

This paper explores the dynamic assortment optimization problem in the context of live-streaming sales. In response to the dynamic characteristics of live-streaming e-commerce, we propose a new choice model that extends the traditional Multinomial Logit model into the continuous time domain. This novel model enables the decoupling of parameter estimation for different products and facilitates the convenient updating of posterior parameters at any time. Finally, we introduce a myopic optimization algorithm based on Thompson Sampling. This algorithm effectively balances exploration and exploitation, captures fluctuations in the number of audiences, and demonstrates superior numerical performance.

1 INTRODUCTION

Assortment optimization problems are prevalent across numerous industries, such as retailing, e-commerce, and online advertising. In these contexts, sellers are tasked with choosing a subset from a pool of alternative items, all with the aim of maximizing expected revenue. For a comprehensive understanding of assortment optimization problems, we direct readers to the survey paper of Kök et al. (2015). Notably, dynamic assortment optimization constitutes a critical segment within this domain (Sauré and Zeevi 2013; Miao and Chao 2021). In dynamic settings, customers' preferences are unknown, and thus sellers are confronted with the dual challenge of deciphering customers' preferences while striving to maximize revenue. Given the inherent uncertainty in preference learning, achieving a suitable balance between exploration and exploitation becomes essential. As a result, dynamic assortment problem can be aptly framed as a Multi-Armed Bandit (MAB) problem (Auer et al. 2002). Under this framework, various techniques, including Upper Confidence Bound (UCB) and Thompson Sampling (TS), have been thoroughly investigated (Agrawal et al. 2017; Agrawal et al. 2019).

In this paper, our focus is on an emerging business model that is garnering increasing attention from researchers (Simchi-Levi et al. 2021), yet has not received adequate exploration in the assortment literature: live-streaming sales. Generally, live-streaming sales events are led by key opinion leaders (KOLs) on platforms like Amazon Live, Taobao Live, and TikTok. During the live-streaming sales, hosts meticulously curate products to showcase, promote, and sell successively. The process of product selection emerges as a crucial aspect of a company's preparations for live-streaming sales.

Unlike traditional e-commerce platform retailing, live-streaming sales exhibit distinctive characteristics that existing modeling and optimization methods have not fully addressed. First, conventional literature typically assumes that sellers display a set of multiple products simultaneously. However, in live-streaming sales, products are introduced sequentially, leading to a gradual expansion of the assortment available to customers over time. This dynamic resembles the monotonic choice model proposed by Davis et al. (2015). Consequently, besides the subset selection, sellers must further determine the order and timing of introducing new products, significantly augmenting the degree of freedom in assortment policy. Second, traditional choice models primarily rely on the Multinomial Logit (MNL) model or its variations (Rusmevichientong et al. 2014; Bai et al. 2024), typically assuming that customers make decisions at given discrete time points based on their preferences. However, during live-streaming sales, customers have the flexibility to enter or exit the live-streaming room and make purchase decisions at any moment along a continuous timeline. Therefore, the fluctuation in audience traffic over time emerges as a crucial factor for revenue optimization, which has been scarcely addressed in existing literature. To capture these distinctive characteristics, we propose a novel basic model to formulate the assortment optimization problem in the context of live-streaming. In addressing this problem, our study derives the associated parameter estimation method and develops a Thompson Sampling-based myopic assortment policy.

The rest of this paper is organized as follows. In Section 2, we present the model formulation for live-streaming assortment problem, highlighting its connection to traditional MNL models. Section 3 derives the posterior parameters in the newly introduced model. In Section 4, we propose a myopic algorithm based on Thompson Sampling (TS) to approximately solve the live-streaming optimization problem. Section 5 tests the empirical performance of our proposed algorithm.

2 PROBLEM FORMULATION

In this section, we formulate the assortment optimization problem in the context of live-streaming. First, we present the primary notations and definitions for modeling the live-streaming process and its associated decision-making framework. Then we compare the new basic model with the MNL model, clarifying their distinctions and interrelations.

2.1 Basic Model

For convenience, we denote the set of all candidate products by $\mathscr{M} \triangleq \{1,2,...,M\}$. A live-streaming sales event is denoted by $S(t),\ t \in [0,T]$, where $S(t)$ is the product being introduced by the live-streaming host at time t. The total assortment size for the live-streaming, denoted by $k(T) = \left|\bigcup_{t\in[0,T]} S(t)\right|$, is not predetermined. During the live-streaming, the host starts to introduce the i-th product $s_i \in \mathscr{M}$ at time T_i and continues until time T_{i+1}, where $0 = T_1 < T_2 < \cdots < T_{k(T)} < T_{k(T)+1} = T$. Moreover, the introduction time $T_{i+1} - T_i,\ i = 1,\dots,k(T)$ is constrained to be within $[\Delta T_{\min}, \Delta T_{\max}]$, where $\Delta T_{\min}$ and $\Delta T_{\max}$ are predetermined by decision-makers. For $t \in [T_i, T_{i+1}),\ i = 1,\dots,k(T)$, we have $S(t) = \{s_i\}$. Decision-makers need to determine when to introduce a new product (i.e., $T_i,\ i = 1,\dots,k(T)$) and which new product to introduce (i.e., $s_i,\ i = 1,\dots,k(T)$), where $k(T)$ is also a decision variable to be determined. Consequently, a live-streaming assortment policy is denoted by a sequence $\pi_{[0,T]} = \{(T_i^\pi, s_i^\pi)\}_{i=1}^{k^\pi(T)}$. Clearly the realization of $S(t)$ depends on $\pi_{[0,T]}$, but for the sake of notation simplicity, we omit this dependency in our notation.

Consider an audience entering the live-streaming room at time t_0. The audience will watch the live-streaming for a while and then leave the live-streaming room either after purchasing one product or without purchasing. Before leaving, the audience has access to the assortment set $A(t_0,t) = \bigcup_{t'\in[t_0,t]} S(t')$ at time t. We introduce a set of random variables $\{Y_j, j \in \{0\} \cup A(t_0,t)\}$, where $Y_j,\ j \in A(t_0,t)$ represents the time when the audience's desire to purchase product j matures, and Y_0 represents the time when the audience's desire not to purchase anything matures. Naturally, $Y_j,\ j \in A(t_0,t)$ must not be earlier than the time (TS_j) when the audience first encounters the product j, while Y_0 must not be earlier than the time (TS_0) when the audience enters the live-streaming room. Regardless of which desire matures first, the audience will leave the live-streaming room subjectively, where "subjectively" means the live-streaming will continue as long as possible without forcing the audience to leave, i.e., $T = +\infty$ mathematically. Conditional on the assortment set $A(t_0,t'),\ t' \geq t$ is fixed as $A(t_0,t)$, the time that the audience leaves the live-streaming room subjectively is denoted by $X|A(t_0,t)$ and supported on $[t,+\infty)$, following a distribution $X|A(t_0,t) \sim \min\{Y_j | Y_j \geq t,\ j \in \{0\} \cup A(t_0,t)\}$.

At $t = 0$, the live-streaming begins with an initial audience count of n_0. The arrival process of new audiences is known and denoted by $N(t),\ t \in [0,T]$, where $N(t)$ is the flow intensity of audiences entering the live-streaming room at time t. During the live-streaming, each product $j \in \mathscr{M}$ yields a known reward of r_j upon sale. At $t = T$, the live-streaming ends, and all remaining audiences are subsequently forced to leave, with no further purchasing actions occurring. The expected total reward under policy $\pi_{[0,T]}$ is denoted by $R(\pi_{[0,T]})$, and our objective is to design an optimal policy that maximizes $R(\pi_{[0,T]})$, i.e.,

$$\max_{\pi_{[0,T]}} R(\pi_{[0,T]}). \tag{1}$$

2.2 Connection to MNL Model

The density function of Y_j's is generally flexible, allowing our model to capture the underlying purchasing behavior of the audience effectively. For the purpose of establishing a connection to the classic MNL model, we assume in this study that $Y_j - TS_j$ follows an exponential distribution (Exp(λ_j)) with a rate parameter λ_j, $j \in \{0\} \cup \mathscr{M}$ and is independent of each other.

First, we provide a brief overview of the popular MNL model. In the MNL model, v_j represents the preference weight that customers assign to product j, while v_0 denotes the preference weight associated with the no-purchase option. Given that we offer the set of products A to customers, each customer purchases product $j \in A$ with a probability of $v_j/(v_0 + \sum_{\ell \in A} v_\ell)$. Within the MNL model framework, customers observe an assortment and make a purchasing choice instantaneously. Thus, the MNL model can be regarded as a static choice model for discrete-time points.

In our study, the purchasing behavior of audiences is modeled as a stochastic process. The audience continuously observes the assortment and makes a purchasing decision at the conclusion of this stochastic process. We further investigate the probability of purchasing product j conditional on the audience leaves at time $t' \in [T_i, T_{i+1}]$ and given $A(t_0, T_i)$. With the assortment set $A(t_0, t)$, $t \in [T_i, T_{i+1}]$ being fixed and utilizing the memoryless property, we have when $t' \in [T_i, T_{i+1}]$,

$$(t' - T_i)|A(t_0, T_i) \sim \min\{\mathrm{Exp}(\lambda_j), j \in \{0\} \cup A(t_0, T_i)\} = \mathrm{Exp}\left(\lambda_0 + \sum_{m \in A(t_0, T_i)} \lambda_m\right). \tag{2}$$

Therefore, for x supported on $\{0\} \cup A(t_0, T_i)$ and $t' \in [T_i, T_{i+1}]$, we have

$$\begin{aligned}
\mathbb{P}(x = j|t', A(t_0, T_i)) &= \mathbb{P}(x = j|t' - T_i, A(t_0, T_i)) \\
&= \frac{f(x = j, t' - T_i, A(t_0, T_i))}{f(t' - T_i|A(t_0, T_i))} \\
&= \frac{\lambda_j e^{-(\lambda_0 + \sum_{m \in A(t_0, T_i)} \lambda_m)(t' - T_i)}}{(\lambda_0 + \sum_{m \in A(t_0, T_i)} \lambda_m) e^{-(\lambda_0 + \sum_{m \in A(t_0, T_i)} \lambda_m)(t' - T_i)}} \\
&= \frac{\lambda_j}{\lambda_0 + \sum_{m \in A(t_0, T_i)} \lambda_m},
\end{aligned} \tag{3}$$

where $f(\cdot)$ denotes the probability density function. The choice model in (3) aligns with the MNL model: conditional on the time of making a purchasing choice and the current assortment set, the purchasing choice is multinomially distributed. Meanwhile the rate parameter λ_j can be considered as the preference weight of product j in the MNL model. It is worth noting that our choice model $\mathbb{P}(x = j|t', A(t_0, T_i))$ extends the discrete-time MNL model into the continuous-time domain. Specifically, as shown in Figure 1, a random variable t' is first generated as the departure time, and then the purchasing choice follows an MNL model with respect to the current assortment.

3 PARAMETER ESTIMATION

In this study, we estimate each λ_j, $j \in \{0\} \cup \mathscr{M}$ under a Bayesian framework. Leveraging the conjugacy of exponential families, we assume each λ_j, $j \in \{0\} \cup \mathscr{M}$ follows a prior Gamma(α_j, β_j). At any time t, the available historical information up to t is $H_t = \{\{S(t')\}_{t' \in [0,t]}, \{\mathbf{c}_n(t), TE_n, TL_n(t)\}_{n=1,\ldots,n_t}\}$,

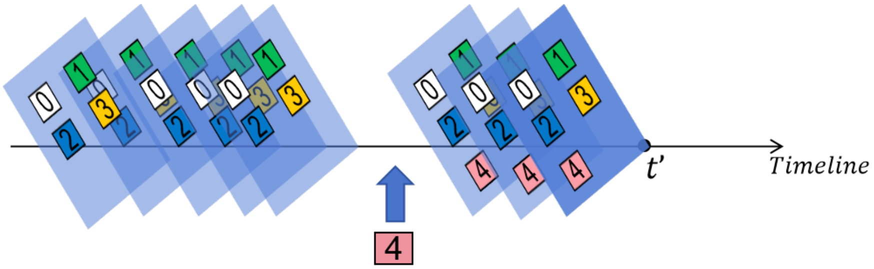

Figure 1: Audiences may make decisions at any moment t' along a continuous timeline, following the MNL model.

where $n_t = n_0 + \int_0^t N(x)dx$ is the total number of audiences entering the live-streaming room before t, $\mathbf{c}_n(t) \triangleq (c_{n,0}(t), c_{n,1}(t), \dots, c_{n,M}(t))$ is a vector with each of its dimension $c_{n,j}(t)$ indicating whether audience n purchases product j or not until time t, TE_n is the entering time of audience n, and $TL_n(t)$ is the latest watching time of audience n. Given H_t, we have

$$k(t) = \left| \bigcup_{t' \in [0,t]} S(t') \right|,$$

$$T_i = \inf\{t' | k(t') = i, t' \in [0,t]\},$$

$$s_i = S(T_i),$$

$$d_{n,i}(t) = \begin{cases} T_{i+1} - T_i, & TE_n < T_i \text{ and } TL_n(t) > T_{i+1} \\ TL_n(t) - T_i, & TE_n < T_i < TL_n(t) \le T_{i+1} \\ T_{i+1} - TE_n, & T_i \le TE_n < T_{i+1} < TL_n(t) \\ TL_n(t) - TE_n, & T_i \le TE_n < TL_n(t) \le T_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

where $k(t)$ is the assortment size for the live-streaming at time t, $d_{n,i}(t)$ is the time duration of audience n watching product s_i up to time t, and then $\mathbf{d}_n(t) \triangleq (d_{n,1}(t), \dots, d_{n,k(t)}(t))$. Therefore, the historical information set H_t can be reformulated as $H_t = \{k(t), \{T_i, s_i\}_{i=1,\dots,k(t)}, \{\mathbf{c}_n(t), \mathbf{d}_n(t)\}_{n=1,\dots,n_t}\}$.

By Bayes' rule, the posterior joint distribution of $\boldsymbol{\lambda} \triangleq (\lambda_0, \lambda_1, \dots, \lambda_M)$ conditional on H_t is derived as follows:

$$f(\boldsymbol{\lambda}|H_t) \propto f(H_t|\boldsymbol{\lambda}) f(\boldsymbol{\lambda}).$$

As λ_j's are independent with each other, we have the prior distribution

$$f(\boldsymbol{\lambda}) = \prod_{j \in \{0\} \cup \mathscr{M}} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} e^{-\beta_j \lambda_j}.$$

As all audiences are independent with each other, we have the likelihood function

$$f(H_t|\boldsymbol{\lambda}) = \prod_{n=1}^{n_t} f(k(t), \{T_i, s_i\}_{i=1,\dots,k(t)}, \mathbf{c}_n(t), \mathbf{d}_n(t) | \boldsymbol{\lambda}).$$

Derived from (2), for each audience n, we have

$$\begin{aligned}
& f(k(t),\{T_i,s_i\}_{i=1,\dots,k(t)},\mathbf{c}_n(t),\mathbf{d}_n(t)|\boldsymbol{\lambda}) \\
= & \left(\prod_{i=1}^{k(t)-1}\left[\left(\lambda_0 c_{n,0}(t)+\sum_{i'=1}^{i}\mathbb{1}\{d_{n,i'}(t)>0\}\lambda_{s_{i'}}c_{n,s_{i'}}(t)\right)^{1-\mathbb{1}\{d_{n,i+1}(t)>0\}}\right.\right. \\
& \left.\left. e^{-\left(\lambda_0+\sum_{i'=1}^{i}\mathbb{1}\{d_{n,i'}(t)>0\}\lambda_{s_{i'}}\right)d_{n,i}(t)}\right]^{\mathbb{1}\{d_{n,i}(t)>0\}}\right)\times\left[\left(\lambda_0 c_{n,0}(t)+\sum_{i=1}^{k(t)}\mathbb{1}\{d_{n,i}(t)>0\}\lambda_{s_i}c_{n,s_i}(t)\right)^{\sum_{i=0}^{k(t)}c_{n,i}(t)}\right. \\
& \left. e^{-\left(\lambda_0+\sum_{i=1}^{k(t)}\mathbb{1}\{d_{n,i}(t)>0\}\lambda_{s_i}\right)d_{n,k(t)}(t)}\right]^{\mathbb{1}\{d_{n,k(t)}(t)>0\}} \\
= & \left(\lambda_0^{c_{n,0}(t)}e^{-\sum_{i=1}^{k(t)}\lambda_0 d_{n,i}(t)}\right)\times\left(\prod_{i=1}^{k(t)}\lambda_{s_i}^{c_{n,i}(t)}e^{-\sum_{i'=i}^{k(t)}\lambda_{s_i}d_{n,i'}(t)}\right).
\end{aligned}$$

Therefore, the posterior distribution of $\lambda_j,\ j\in\{0\}\cup\mathscr{M}$ is shown as follows:

$$\lambda_j|H_t\sim\begin{cases}\text{Gamma}(\alpha_j+\sum_{n=1}^{n_t}c_{n,j}(t),\beta_j+\sum_{n=1}^{n_t}\sum_{i=1}^{k(t)}d_{n,i}(t)), & j=0\\ \text{Gamma}(\alpha_j+\sum_{n=1}^{n_t}c_{n,j}(t),\beta_j+\sum_{n=1}^{n_t}\sum_{i'=i}^{k(t)}d_{n,i'}(t)), & j=s_i,\ i=1,\dots,k(t)\\ \text{Gamma}(\alpha_j,\beta_j), & \text{otherwise.}\end{cases}$$

Note that, even without purchasing product j, we can still update the posterior parameter of product j using non-purchase data and data from other products. Additionally, the posterior updates of different products are decoupled, which cannot be done in traditional MNL models.

4 THOMPSON SAMPLING-BASED MYOPIC POLICY

The seller's strategy space encompasses both the timing for introducing new products and the selection of which products to introduce. Additionally, the posterior distribution of parameters is updated online during the live-streaming. Due to the vast strategic space and the dynamic nature of the problem, we focus on providing an efficient myopic algorithm to approximately solve the optimization problem (1). Here, "myopic" refers to: (i) looking ahead and optimizing the reward within the subsequent time interval where the product assortment remains unchanged, and (ii) focusing solely on the profits brought by the current customers in the live-streaming, disregarding potential profits from future customers. The utilization of the myopic algorithm to approximately solve the original optimization problem is a prevalent practice within the assortment optimization literature, particularly for online and dynamic settings (Simchi-Levi et al. 2021; Gong et al. 2022; Aouad and Saban 2023).

For the sake of theorem formulation convenience, we denote the expected reward of audience n leaving the live-streaming at time t by $R_n(t)|A_n(t)$, where $A_n(t)$ represents the assortment set for audience n at time t. Thus following (3), $R_n(t)|A_n(t)=(\sum_{m\in A_n(t)}r_m\lambda_m)/(\lambda_0+\sum_{m\in A_n(t)}\lambda_m)$. Further, if the assortment set is fixed as $A_n(t)$ during the period $[t,t+\Delta T]$, then following (2) and (3), the expected "myopic" reward of audience n during the period $[t,t+\Delta T]$ can be denoted and calculated by

$$R_n([t,t+\Delta T])|A_n(t)=\frac{\sum_{m\in A_n(t)}r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)}\lambda_m}\left(1-e^{-(\lambda_0+\sum_{m\in A_n(t)}\lambda_m)\Delta T}\right). \tag{4}$$

First, we discuss the strategy for the timing of product introductions. To begin with, we present the following theorem to demonstrate an optimality property of the assortment policy.

Theorem 1 Given any period $[t,t+\Delta T]$ such that $\left|\bigcup_{t'\in[t,t+\Delta T]} S(t')\right| = 1$, let N_t be the number of audiences at time t and $\bar{R}(t) \triangleq \max\limits_{n=1,\dots,N_t} R_n(t)|A_n(t)$. If there is a product j that has not yet been introduced such that $r_j \geq \bar{R}(t)$, then assortments $A_n(t)\cup\{j\},\ n=1,\dots,N_t$ has a better expected reward of audience n during the period $[t,t+\Delta T]$ than assortments $A_n(t),\ n=1,\dots,N_t$.

Proof. Given the assortment $A_n(t)$, the expected reward of audience n during the period $[t,t+\Delta T]$ is

$$R_n([t,t+\Delta T])|A_n(t) = \frac{\sum_{m\in A_n(t)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)}\lambda_m}\left(1-e^{-(\lambda_0+\sum_{m\in A_n(t)}\lambda_m)\Delta T}\right).$$

Given the assortment $A_n\cup\{j\}$, the expected reward of audience n during the period $[t,t+\Delta T]$ is

$$R_n([t,t+\Delta T])|A_n(t)\cup\{j\} = \frac{\sum_{m\in A_n(t)\cup\{j\}} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)\cup\{j\}}\lambda_m}\left(1-e^{-(\lambda_0+\sum_{m\in A_n(t)\cup\{j\}}\lambda_m)\Delta T}\right).$$

Given the sufficient condition

$$r_j \geq \max_{n=1,\dots,N_t}\frac{\sum_{m\in A_n(t)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)}\lambda_m} \geq \frac{\sum_{m\in A_n(t)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)}\lambda_m},$$

we have

$$\frac{\sum_{m\in A_n(t)\cup\{j\}} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)\cup\{j\}}\lambda_m} \geq \frac{\sum_{m\in A_n(t)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(t)}\lambda_m}.$$

And it is obvious that

$$\left(1-e^{-(\lambda_0+\sum_{m\in A_n(t)\cup\{j\}}\lambda_m)\Delta T}\right) > \left(1-e^{-(\lambda_0+\sum_{m\in A_n(t)}\lambda_m)\Delta T}\right),$$

then we have $R_n([t,t+\Delta T])|A_n(t)\cup\{j\} > R_n([t,t+\Delta T])|A_n(t)$, which concludes this theorem. □

As implied by Theorem 1, the quantity $\bar{R}(t)$ is an important indicator for determining the timing of product introductions. Let $\bar{r}(t) = \max_{j\notin\bigcup_{t'\in[0,t)}S(t')} r_j$, then the *profitable condition*

$$\bar{R}(t) < \bar{r}(t) \tag{5}$$

is a sufficient condition for the existence of a product whose addition can lead to higher "myopic" rewards in the subsequent time period. In our policy, the *profitable condition* at time t serves as the criterion for deciding whether to introduce a new product into the assortment collection at that moment. Specifically, as shown in Algorithm 1, the timing policy entails the following: once a product has been presented for at least the minimum explanation time $\Delta T_{\min}$, the seller repeatedly checks the *profitable condition*, and as soon as this condition is met, the next product can be introduced. When computing $\bar{R}(t)$ and $\bar{r}(t)$ in practice, the parameters $\lambda_j,\ j\in\mathscr{M}\cup\{0\}$ are all replaced by their posterior estimates.

In general, the entire live-streaming process can be divided into two phases: the *growth phase* and the *stable phase*, which are separated by a hitting time defined as

$$\tau = \inf\{t\geq 0 : \bar{R}(t)\geq\bar{r}(t)\}.$$

Here, τ is the time when the *profitable condition* is first violated, and consequently, the *growth phase* refers to the interval $[0,\tau)$, while the *stable phase* refers to $(\tau,T]$. Next, we will discuss the assortment selection strategies for the two phases separately. Before proceeding, we introduce additional notations. For a function f of time, $f(t^-)$ and $f(t^+)$ respectively denote the left and right limits of f at t, or more intuitively, the moment just before and just after time t.

Algorithm 1 Thompson Sampling-Based Myopic Policy (TS-Myopic)

```
Input: $\{\alpha_j, \beta_j, r_j\}_{j \in \mathscr{M} \cup \{0\}}$, T, $\Delta T_{\min}$, $\Delta T_{\max}$
Initialization. Randomly introduce a product $s_1 \in \mathscr{M}$ at time $T_1 = 0$. Let $i = 1$ and $\mathbf{I}(t \geq \tau) = 0$.
While $t < T$ do
Posterior Update and Check Profitable Condition. Update $\lambda_j | H_t$ with equation (3). Calculate $\bar{R}(t)$ and $\bar{r}(t)$ using posterior parameter estimates.
If $\bar{R}(t)$ first exceeds $\bar{r}(t)$ then
  Let $\mathbf{I}(t \geq \tau) = 1$.
End if
Product Introduction. If $\mathbf{I}(t \geq \tau) = 0$, go to (A); If $\mathbf{I}(t \geq \tau) = 1$, go to (B).
(A) If $t \geq T_i + \Delta T_{\min}$ then
    Let $i = i + 1$ and $T_i = t$. Introduce the product $s_i$ according to (7) at time $T_i$.
    End if
(B) If "$\bar{R}(t) < \bar{r}(t)$ and $t \geq T_i + \Delta T_{\min}$" or "$t \geq T_i + \Delta T_{\max}$"
    Let $i = i + 1$ and $T_i = t$. Introduce the product $s_i$ according to (8) at time $T_i$.
    End if
End while
```

4.1 Growth Phase

As shown in Figure 2, at the initiation of the live-streaming, the assortment collection contains only one product, s_1, and $\bar{R}(t)$ is significantly lower than $\bar{r}(t)$, $t \in [0, \tau)$, primarily due to λ_0 typically surpassing λ_{s_1}. Consequently, there is a high probability that audiences will choose to leave without purchasing. Therefore, during the *growth phase*, it is crucial to quickly introduce more profitable products to retain customers and compete against the intention to leave without making a purchase. Following the introduction of s_i at time T_i, sellers tend to swiftly introduce subsequent products after the minimal time interval $\Delta T_{\min}$. When $\Delta T_{\min}$ is small enough, utilizing (4) and Taylor expansion, the total expected reward of audiences $n = 1, \cdots, N_{T_i}$ during time $[T_i, T_i + \Delta T_{\min})$ conditioned on assortments $A_n(T_i^+) = A_n(T_i^-) \cup \{s_i\}$, $n = 1, \cdots, N_{T_i}$ can be expressed as

$$\sum_{n=1}^{N_{T_i}} R([T_i, T_i + \Delta T_{\min})) | A_n(T_i^+) = \sum_{n=1}^{N_{T_i}} \sum_{m \in A_n(T_i^+)} r_m \lambda_m \Delta T_{\min} + o(\Delta T_{\min}). \tag{6}$$

It is straightforward to verify that selecting a product s_i maximizing the "myopic" reward in (6) is equivalent to maximizing $r_{s_i} \lambda_{s_i}$. Therefore, products with higher $r_j \lambda_j$ are more preferable. The most straightforward idea is to pick the product that maximizes the posterior mean of $r_j \lambda_j$. However, this strategy, being overly greedy, lacks exploration, particularly in scenarios where parameter estimates lack sufficient accuracy. To address this, we turn to MAB techniques such as UCB and TS to strike a balance between exploration and exploitation. Given the computational simplicity of the proposed model's posterior updates and sampling, we opt for TS. From the MAB perspective, introducing a new product j is akin to pulling the "arm" j, with the resulting reward being $r_j \lambda_j$. In our algorithm, parameters are sampled independently from the current posterior distribution, denoted as $\hat{\lambda}_1, \cdots, \hat{\lambda}_M$, where $\hat{\lambda}_j \sim \lambda_j | H_{T_i}$, $j \in \mathscr{M}$. In the growth phase, the product introduced at time T_i is chosen as

$$s_i = \operatorname{argmax}_{j \notin \bigcup_{t' \in [0, T_i)} S(t')} \quad r_j \hat{\lambda}_j. \tag{7}$$

4.2 Stable Phase

As the live-streaming progresses and new profitable products are continually introduced, generally, $R_n(t) | A_n(t)$ as well as $\bar{R}(t)$ increase over time, while the maximum reward of the remaining products,

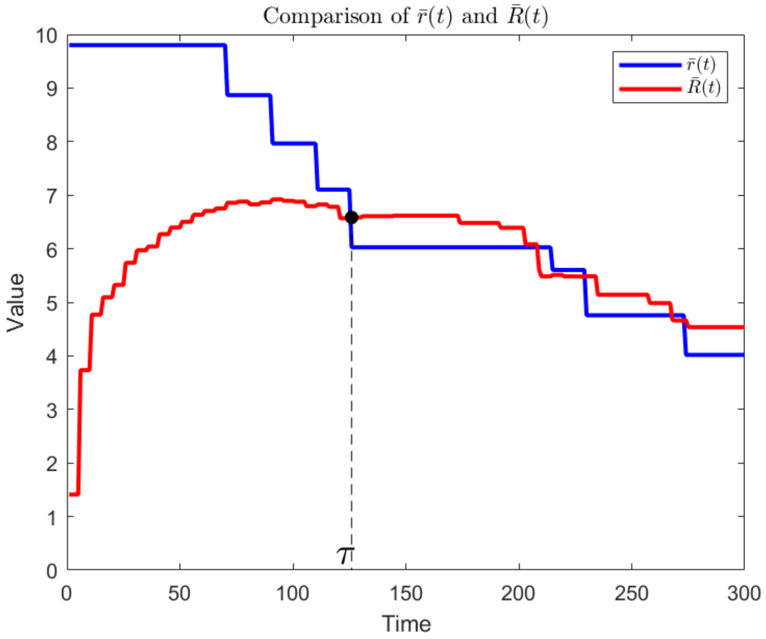

Figure 2: An illustration of profitable condition and hitting time τ.

$\bar{r}(t)$, decreases over time. Eventually, at time τ, $\bar{R}(t)$ surpasses $\bar{r}(t)$, marking the onset of the stable phase. At this point, introducing new products may not increase profits. Sellers should wait until those audiences with high $R_n(t)|A_n(t)$ have left the live-streaming room, causing $\bar{R}(t)$ to drop back below $\bar{r}(t)$, before considering the introduction of new products. Assuming that after product s_i is introduced at time T_i, condition (5) is violated. Let $\mathscr{N}^*(T_i) = \{n \in \{1,\cdots,N_t\} \Big| R_n(T_i)|A_n(T_i) = \bar{R}(T_i)\}$ be the set of audiences with the maximum $R_n|A_n$ at time T_i, and note that all audiences in this set share the same available assortment $A^*(T_i)$. $\bar{R}(t)$ decreases if and only if all customers belonging to $\mathscr{N}^*(T_i)$ have left. Thus, by calculating the expected maximum duration of customers in the live-streaming room, we obtain the lower bound of the expected time until the next product introduction as

$$\mathbb{E}(\Delta T) \geq \left(1+\frac{1}{2}+\cdots+\frac{1}{|\mathscr{N}^*(T_i)|}\right)\sum_{j\in A^*(T_i)}\lambda_j \sim \mathscr{O}\big(\log(|\mathscr{N}^*(T_i)|)\big),$$

which signifies the time required for all audiences in $\mathscr{N}^*(T_i)$ to leave.

Therefore, in the stable phase, we assume that ΔT is relatively large and with (4), we approximate the "myopic" reward during $[T_i, T_i+\Delta T)$ as

$$\begin{aligned}
&\sum_{n=1}^{N_{T_i}} R([T_i, T_i+\Delta T))|A_n(T_i^+)\\
&=\sum_{n=1}^{N_{T_i}}\left(\frac{\sum_{m\in A_n(T_i^+)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(T_i^+)}\lambda_m}+\mathscr{O}(e^{-(\lambda_0+\sum_{m\in A_n(T_i^+)}\lambda_m)\Delta T})\right)\\
&\approx\sum_{n=1}^{N_{T_i}}\frac{\sum_{m\in A_n(T_i^+)} r_m\lambda_m}{\lambda_0+\sum_{m\in A_n(T_i^+)}\lambda_m},
\end{aligned}$$

which exactly equals $\sum_{n=1}^{N_{T_i}} R_n(T_i^+)|A_n(T_i^+)$. The following corollary of Theorem 1 suggests that in the stable phase, introducing products with higher λ is preferable. The technique of its proof is similar to that of Theorem 1 and is omitted.

Corollary 2 Suppose that product s_i is added to the assortment at time T_i, i.e. $A_n(T_i^+) = A_n(T_i^-) \cup \{s_i\}$ and the condition $r_{s_i} > \bar{R}(T_i)$ holds. Then for any $n = 1, \cdots, N_{T_i}$, $R_n(T_i^+)|A_n(T_i^+)$ is increasing with respect to λ_{s_i}.

In our TS-based algorithm, the parameters of products that are introduced before T_i are estimated by their posterior mean, i.e.,

$$\hat{\lambda}_j = \mathbf{E}(\lambda_j | H_{T_i}), \quad j \in \left(\bigcup_{t' \in [0,T_i)} S(t') \right) \bigcup \{0\}.$$

The parameters $\hat{\lambda}_j$, $j \notin \bigcup_{t' \in [0,T_i)} S(t')$ of other products that are not introduced before are sampled from their posterior distribution $\lambda_j | H_{T_i}$ independently. In the stable phase, the product introduced at time T_i is chosen as

$$s_i = \text{argmax}_{j \notin \bigcup_{t' \in [0,T_i)} S(t')} \sum_{n=1}^{N_{T_i}} \frac{\sum_{m \in A_n(T_i^+)} r_m \hat{\lambda}_m}{\hat{\lambda}_0 + \sum_{m \in A_n(T_i^+)} \hat{\lambda}_m}, \tag{8}$$

where $A_n(T_i^+) = A_n(T_i^-) \cup \{j\}$.

5 NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of the newly proposed TS-based myopic (TS-Myopic) policy through simulation experiments. In this experiment, the total number of candidate products is $M = 60$, with a total live-streaming time of $T = 300$. The parameters of each product are $\lambda_j = 500(15 + \frac{5j}{12})$, where $j = 1, \cdots, 60$ and r_j are generated from uniform distribution $U[0,10]$. The parameter of "no-purchase" is $\lambda_0 = 5000$. The minimum and maximum introduction times for each product are $\Delta T_{\min} = 5$ and $\Delta T_{\max} = 50$, respectively. The arrival of customers follows a Poisson distribution with an arrival rate of 10. The initial number of customers at the start of the live-streaming process is $n_0 = 100$.

One of the benchmarks in the experiment is the static monotonic policy (referred to as "Static") introduced in Davis et al. (2015). This policy is built upon a scenario where the assortment size grows over time, akin to the live-streaming setting. However, it relies on static optimization and maintains fixed intervals for the introduction of products, ignoring the dynamics of customer behavior. Another benchmark is the completely random policy (referred to as "Random"), where, given a total assortment size K, a random product is introduced at random time. We provide two versions of random policy: one with $K = 20$ and another with $K = 40$. Moreover, we consider two scenarios: known and unknown parameters. In the known parameters case, all λ parameters in the TS-Myopic and Static policy are replaced by their true values rather than using TS samples or posterior estimates. In the unknown parameters case, the prior distribution of λ_j is set to be Gamma(α_j, β_j), with $\alpha_j = 100(\lambda_j + 500\varepsilon_j)$, where ε_j is generated from normal distribution $N(0,1)$, and $\beta_j = 100$. The Static algorithm updates the posterior parameters after each new product addition and then updates the assortment policy. Additionally, it is worth mentioning that the performance of the random policy is unaffected by whether the parameters are known or not.

The experimental results, illustrating the total reward accrued over time, are showcased in Figure 3. Regardless of whether the parameters are known or not, the performance of the TS-Myopic algorithm surpasses that of all other benchmarks. Before time $t < 150$, due to parameter estimation bias, the TS-Myopic algorithm with unknown parameters exhibits lower performance compared to its counterpart with known parameters. However, after $t > 150$, as parameter estimation becomes increasingly accurate, the performance of the TS-Myopic policy in both cases becomes nearly identical, resulting in a parallel growth of cumulative rewards. Moreover, the Static policy proves effective in addressing static monotonic assortment optimization challenge. Nonetheless, surprisingly, even in the case where parameters are known, the performance of the Static algorithm is inferior to that of the Random algorithm with $K = 40$. This suggests that in the context of live-streaming, the dynamics resulting from audience influx and outflow is the primary factor influencing rewards management.

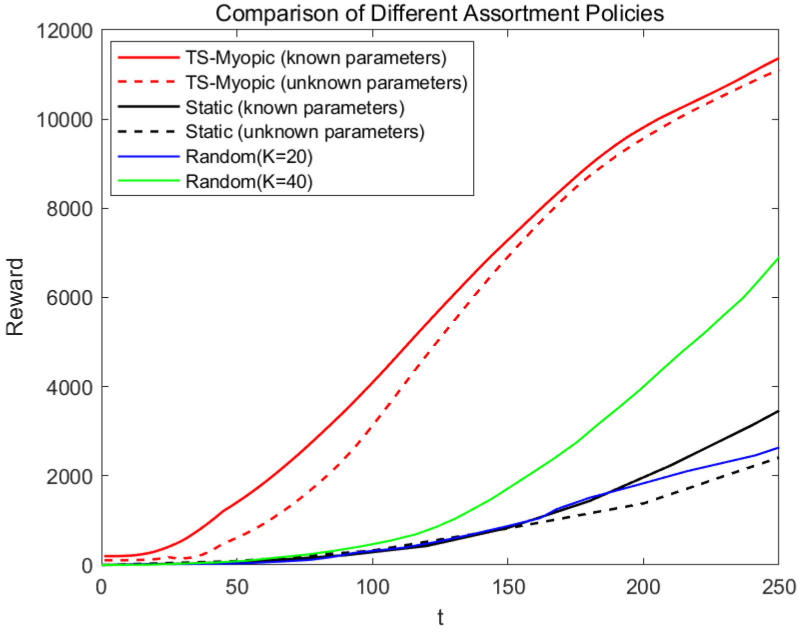

Figure 3: Evolution of cumulative reward over time t.

6 CONCLUSION

The unique attributes of live-streaming e-commerce pose significant challenges to conventional assortment problems discussed in literature, particularly concerning modeling, parameter estimation, and optimization algorithms. Our newly introduced model extends the traditional choice model into the continuous time domain. This model also allows for the convenient updating of parameter estimates at any moment. Furthermore, in comparison to traditional static assortment optimization methods, our proposed TS-based myopic algorithm effectively captures the dynamic nature of live-streaming sales, yielding superior empirical outcomes. Additionally, our continuous-time choice model can be applied not only in live-streaming sales but also in the assortment problems of traditional e-commerce platforms.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China (NSFC) under Grants 72192843, 72201006.

REFERENCES

Agrawal, S., V. Avadhanula, V. Goyal, and A. Zeevi. 2017. "Thompson Sampling for the MNL-Bandit". In *Proceedings of the 2017 Conference on Learning Theory*, edited by S. Kale and O. Shamir, 76–78. New York: Proceedings of Machine Learning Research.

Agrawal, S., V. Avadhanula, V. Goyal, and A. Zeevi. 2019. "MNL-bandit: A Dynamic Learning Approach to Assortment Selection". *Operations Research* 67(5):1453–1485.

Aouad, A. and D. Saban. 2023. "Online Assortment Optimization for Two-sided Matching Platforms". *Management Science* 69(4):2069–2087.

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. “Finite-time Analysis of the Multiarmed Bandit Problem”. *Machine Learning* 47(2–3):235–256.

Bai, Y., J. Feldman, H. Topaloglu, and L. Wagner. 2024. “Assortment Optimization under the Multinomial Logit Model with Utility-based Rank Cutoffs”. *Operations Research* 72(4):1453–1474.

Davis, J. M., H. Topaloglu, and D. P. Williamson. 2015. “Assortment Optimization Over Time”. *Operations Research Letters* 43(6):608–611.

Gong, X.-Y., V. Goyal, G. N. Iyengar, D. Simchi-Levi, R. Udwani, and S. Wang. 2022. “Online Assortment Optimization with Reusable Resources”. *Management Science* 68(7):4772–4785.

Kök, A. G., M. L. Fisher, and R. Vaidyanathan. 2015. “Assortment Planning: Review of Literature and Industry Practice”. In *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, edited by N. Agrawal and S. A. Smith, 175–236. Boston, MA: Springer US.

Miao, S. and X. Chao. 2021. “Dynamic Joint Assortment and Pricing Optimization with Demand Learning”. *Manufacturing & Service Operations Management* 23(2):525–545.

Rusmevichientong, P., D. Shmoys, C. Tong, and H. Topaloglu. 2014. “Assortment Optimization under the Multinomial Logit Model with Random Choice Parameters”. *Production and Operations Management* 23(11):2023–2039.

Sauré, D. and A. Zeevi. 2013. “Optimal Dynamic Assortment Planning with Demand Learning”. *Manufacturing & Service Operations Management* 15(3):387–404.

Simchi-Levi, D., Z. Zheng, and F. Zhu. 2021. “Dynamic Planning and Learning under Recovering Rewards”. In *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, 9702–9711. New York: Poceedings of Machine Learning Research.

AUTHOR BIOGRAPHIES

ZISHI ZHANG is a Ph.D. candidate in Guanghua School of Management at Peking University, Beijing, China. His email address is zishizhang@stu.pku.edu.cn.

HAIDONG LI is an Assistant Professor at the School of Economics and Management, University of Chinese Academy of Sciences (UCAS). His research interests include simulation optimization, stochastic estimation, and reinforcement learning. His email address is haidong.li@ucas.ac.cn.

YING LIU is a Professor at the School of Economics and Management and MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences. His research interests include digital economy, data analysis, Fintech and E-commerce. His email address is liuy@ucas.ac.cn.

YIJIE PENG is an Associate Professor in Guanghua School of Management at Peking University, Beijing, China. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare. His email address is pengyijie@pku.edu.cn.