# Distributionally Robust Choice Models

B. Tech Project Report<br>
By<br>
Krishanveni Unnikrishnan [2006321]<br>
Under the guidance of<br>
Dr. Divya Padmanabhan<br>
Spring 2024

School of Mathematics and Computer Science<br>
**Indian Institute of Technology Goa**

# Acknowledgement

I would like to express my sincere gratitude to Dr Divya Padmanabhan for her guidance and constant support throughout the project. I would also take this opportunity to thank Tina Sher, PhD student in the School of Mathematics and Computer Science, IIT Goa and Somesh Aggrawal, fourth year undergraduate student in the School of Mathematics and Computer Science, IIT Goa for their invaluable contributions to the project.

# Contents

# 1 Introduction

I have explored two verticals for this project. Firstly, I studied and replicated the results of the paper “MNL-Bandit: A Dynamic Learning Approach to Assortment Selection”. The paper uses a UCB Multi Arm Bandit strategy to solve the assortment selection problem dynamically. Secondly, I tried to extend a similar Multi Arm Bandit framework to learn the underlying probability distribution of a Marginalized Distribution Model (MDM) Choice Model.

Assortment optimization problems arise widely in many industries including retailing and online advertising where the seller needs to select a subset from a universe of substitutable items with the objective of maximizing expected revenue. In order to capture substitution effects among products we use choice models.

The goal is to devise a policy i.e., a sequence of assortments $S_1.S_2, ...., S_T$ where $T$ is the planning horizon such that the cumulative expected revenues over said horizon is maximized. Alternatively, minimizing the gap between performance of a proposed policy and that of an oracle that knows instance parameters a priori. This gap is called regret.

# 2 Background

## 2.1 Choice Models

Choice models are statistical models used to understand and predict how choices are made between different alternatives where alternatives are described in terms of attributes. For example, if the alternatives are cars, attributes could be brand, price, fuel efficiency, resale value etc. consumers will choose the alternative that provides them with the highest level of utility which is a measure of the desirability or attractiveness of an alternative.

Choice modeling involves experiments where consumers are presented with a set of alternatives and asked to choose their preferred option. Once the data is collected, choice models are estimated using statistical techniques such as logistic regression, multinomial logit, nested logit, or more advanced methods like mixed logit or hierarchical Bayes models. These models estimate the relationship between the attributes of alternatives and the likelihood of them being chosen.

Choice models are built on a probabilistic framework, where the probability of choosing a particular alternative is modeled as a function of the attributes of that alternative and possibly other factors. Through this project we are trying to infer this underlying probability distribution of each product.

## 2.2 Multinomial Choice Models

Some of the multinomial choice models are:

- Multinomial Logit (MNL) is a basic model for modeling choices among multiple alternatives, where each alternative is assigned a probability based on its attributes.
- Nested Logit is an extension of MNL that accounts for correlations among alternatives within predefined groups or nests.
- Mixed Logit allows for random parameters in the utility function, capturing heterogeneity in preferences across individuals.
- Lastly, there is Generalized Extreme Value (GEV) Models. They are a flexible class of models that includes MNL and nested logit as special cases and can accommodate various types of utility functions.

The choice probability in the Multinomial Logit (MNL) model whose attributes are $X$ and parameters of product $i$ is $\beta_i$ is given by:

$$P(i|X) = \frac{e^{\beta_i \cdot X}}{\sum_{j=1}^{J} e^{\beta_j \cdot X}} \tag{1}$$

## 2.3 Multi Arm Bandit Frameowork

In a multi-armed bandit problem, there is a set of K arms, each associated with an unknown reward distribution. The agent must select one arm at each time step and receive a reward from the chosen arm according to its distribution. The goal is to maximize the total reward accumulated over a series of selections.

The challenge is balancing exploration and exploitation. Exploration involves trying different arms to learn about their reward distributions, while exploitation involves selecting the arm that is believed to have the highest expected reward based on the available information. The trade-off arises because exploring unknown arms may lead to discovering arms with high rewards, but exploiting known arms may yield immediate rewards.

Some of the strategies to address the exploration-exploitation dilemma are:

1. **Epsilon-Greedy:** With probability $\epsilon$, choose a random arm (explore); otherwise, choose the arm with the highest estimated reward (exploit).

2. **Upper Confidence Bound (UCB):** Choose the arm with the highest upper confidence bound on its expected reward, balancing between exploiting high-reward arms and exploring uncertain arms.

3. **Thompson Sampling:** Sample arms' reward distributions from their posterior distributions and select the arm with the highest sampled reward (exploit), incorporating uncertainty into the decision-making process.

Regret in a multi-armed bandit problem measures the opportunity cost incurred by not always selecting the arm with the highest expected reward. The cumulative regret is the difference between the total reward obtained by the agent and the total reward that would have been obtained by always selecting the arm with the highest expected reward. The goal of the agent is often to minimize regret over time, ensuring that it learns to make better decisions as it gains more information about the arms' reward distributions.

# 3 MNL Bandit Algorithm

## 3.1 Problem Formulation

At every time instance $t$, the seller selects an assortment $S_t \subset \{1, \ldots, N\}$ and observes the customer purchase decision $c_t \in S_t \cup \{0\}$, where $\{0\}$ denotes the no-purchase alternative, which is always available for the consumer. We assume consumer preferences are modeled using a multinomial logit (MNL) model. Under this model, the probability that a consumer purchases product $i$ at time $t$ when offered an assortment $S_t = S \subset \{1, \ldots, N\}$ is given by,

$$p_i(S) := P\left(c_t = i | S_t = S\right) = \begin{cases} \dfrac{v_i}{v_0 + \sum_{j \in S} v_j}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

for all $t$, where $v_i$ is the *attraction parameter* for product $i$ in the MNL model.

The random variables $\{c_t : t = 1, 2, \ldots\}$ are conditionally independent, namely, $c_t$ conditioned on the event $\{S_t = S\}$ is independent of $c_1, \ldots, c_{t-1}$.

The seller is trying to learn these $v_i$ parameters of a consumer's choice model. From (2), the expected revenue when assortment $S$ is offered and the MNL parameters are denoted by the vector $v$ is given by

$$R(S, v) = E\left[\sum_{i \in S} r_i 1\{c_t = i | S_t = S\}\right] = \sum_{i \in S} \frac{r_i v_i}{1 + \sum_{j \in S} v_j}, \tag{3}$$

where $r_i$ is the revenue obtained when product $i$ is purchased and is known a priori.

**Constraints:**

- **Cardinality constraint:** Upper bound on the number of products that can be offered (display window size).
- **Partition matroid constraint:** Products are partitioned into segments and the retailer can pick at most a specific number of products from each partition.

- **Joint display and assortment constraint:** Decide both assortment and display segment of each assortment. The upper bound on the number of products in each display segment.

In this project I have implemented only cardinality constraints.

## 3.2 Proposed algorithm

We divide the time horizon into epochs, where in each epoch we offer an assortment repeatedly until a no purchase outcome occurs. Let $E_\ell$ denote the set of consecutive time steps in epoch $\ell$. $E_\ell$ contains all time steps after the end of epoch $\ell - 1$, until a no-purchase happens in response to offering $S_\ell$, including the time step at which no-purchase happens.

The length of an epoch $|E_\ell|$ conditioned on $S_\ell$ is a geometric random variable with success probability defined as the probability of no-purchase in $S_\ell$. The total number of epochs $L$ in time $T$ is implicitly defined as the minimum number for which $\sum_{\ell=1}^{L} |E_\ell| \geq T$.

At the end of every epoch $\ell$, estimates for the parameters of MNL are updated, which are used in epoch $\ell + 1$ to choose assortment $S_{\ell+1}$. For any time step $t \in E_\ell$, let $c_t$ denote the consumer's response to $S_\ell$, i.e., $c_t = i$ if the consumer purchased product $i \in S_\ell$, and 0 if no-purchase happened. We define $\hat{v}_{i,\ell}$ as the number of times a product $i$ is purchased in epoch $\ell$,

$$\hat{v}_{i,\ell} := \sum_{t \in E_\ell} 1(c_t = i). \tag{4}$$

For every product $i$ and epoch $\ell \leq L$, we keep track of the set of epochs before $\ell$ that offered an assortment containing product $i$, and the number of such epochs. We denote the set of epochs by $\mathcal{T}_i(\ell)$ and the number of epochs by $T_i(\ell)$. That is,

$$\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\tau\}, \quad T_i(\ell) = |\mathcal{T}_i(\ell)|. \tag{5}$$

Then compute $\bar{v}_{i,\ell}$ as the number of times product $i$ was purchased per epoch,

$$\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in \mathcal{T}_i(\ell)} \hat{v}_{i,\tau}. \tag{6}$$

Using these estimates, we compute the upper confidence bounds, $v_{i,\ell}^{\mathsf{UCB}}$ for $v_i$ as,

$$v_{i,\ell}^{\mathsf{UCB}} := \bar{v}_{i,\ell} + \sqrt{\bar{v}_{i,\ell}\frac{48\log(\sqrt{N}\ell + 1)}{T_i(\ell)}} + \frac{48\log(\sqrt{N}\ell + 1)}{T_i(\ell)}. \tag{7}$$

Upper confidence bounds ensure that the likelihood of identifying the parameter value is sufficiently large. We then offer the optimistic assortment in the next epoch, based on the previous updates as follows,

$$S_{\ell+1} := \underset{S\in\mathcal{S}}{\operatorname{argmax}} \max\left\{R(S,\hat{v}) : \hat{v}_i \le v_{i,\ell}^{\mathsf{UCB}}\right\}, \tag{8}$$

where $R(S, \hat{v})$ is as defined in (3). This will lead us to the following optimization problem,

$$S_{\ell+1} := \underset{S\in\mathcal{S}}{\operatorname{argmax}}\, \tilde{R}_{\ell+1}(S), \tag{9}$$

where $\tilde{R}_{\ell+1}(S)$ is defined as,

$$\tilde{R}_{\ell+1}(S) := \frac{\displaystyle\sum_{i\in S} r_i v_{i,\ell}^{\mathsf{UCB}}}{1 + \displaystyle\sum_{j\in S} v_{j,\ell}^{\mathsf{UCB}}}. \tag{10}$$

## 3.3 Algorithm

**Algorithm 1** Exploration-Exploitation algorithm for MNL Bandit

```
Initialization: $v_{i,0}^{\mathsf{UCB}} = 1$ for all $i = 1, \ldots, N$
$t = 1$ ; $\ell = 1$ keeps track of the time steps and total number of epochs respectively
while $t < T$ do
Compute $S_\ell = \underset{S \in \mathcal{S}}{\operatorname{argmax}} \tilde{R}_\ell(S) = \frac{\sum_{i \in S} r_i v_{i,\ell-1}^{\mathsf{UCB}}}{1 + \sum_{j \in S} v_{j,\ell-1}^{\mathsf{UCB}}}$
Offer assortment $S_\ell$, observe the purchasing decision, $c_t$ of the consumer
    if $c_t = 0$ then
Compute $\hat{v}_{i,\ell} = \sum_{t \in E_\ell} 1(c_t = i)$, no. of consumers who preferred $i$ in epoch $\ell$, for all $i \in S_\ell$
Update $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}$, $T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until $\ell$ that offered product $i$
Update $\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in T_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
Update $v_{i,\ell}^{\mathsf{UCB}} = \bar{v}_{i,\ell} + \sqrt{\bar{v}_{i,\ell} \frac{48 \log(\sqrt{N}\ell + 1)}{T_i(\ell)}} + \frac{48 \log(\sqrt{N}\ell + 1)}{T_i(\ell)}$; $\ell = \ell + 1$
    else
$E_\ell = E_\ell \cup t$, time indices corresponding to epoch $\ell$
    end if
$t = t + 1$
end while=0
```

## 3.4 Upper Confidence Regret Bound

**Assumption 3.1**

1. *The MNL parameter corresponding to any product $i \in \{1, \ldots, N\}$ satisfies $v_i \leq v_0 = 1$.*

2. *The family of assortments $\mathcal{S}$ is such that $S \in \mathcal{S}$ and $Q \subseteq S$ implies that $Q \in \mathcal{S}$.*

The first assumption is equivalent to the 'no purchase option' being the most likely outcome. The second assumption implies that removing a product from a feasible assortment preserves feasibility. This holds for most constraints arising in practice including cardinality, and matroid constraints more generally.

**Theorem 1 (Performance Bounds for Algorithm 2)** *For any instance $v = (v_0, \ldots, v_N)$ of the problem with $N$ products, $r_i \in [0, 1]$ and Assumption 3.1, the regret of the policy given by Algorithm 2 at any time $T$ is bounded as,*

$$Reg_\pi(T, v) \leq C_1 \sqrt{NT \log NT} + C_2 N \log^2 NT,$$

*where $C_1$ and $C_2$ are absolute constants (independent of problem parameters).*

### 3.4.1 Proof Outline

**Confidence intervals.**

The first step in the regret analysis is to prove the following two properties of the estimates $v_{i,\ell}^{UCB}$ computed as in (7) for each product $i$.

1. $v_i$ is bounded by $v_{i,\ell}^{\mathsf{UCB}}$ with high probability, and that as a product is offered an increasing number of times

2. The estimates $v_{i,\ell}^{\mathsf{UCB}}$ converge to the true value with high probability.

Mathematically,

**Lemma 3.1** *For every $\ell = 1, \cdots, L$, we have:*

1. *$v_{i,\ell}^{\mathsf{UCB}} \geq v_i$ with probability at least $1 - \frac{6}{N\ell}$ for all $i = 1, \ldots, N$.*

*2. There exists constants $C_1$ and $C_2$ such that*

$$v_{i,\ell}^{\mathsf{UCB}} - v_i \leq C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell + 1)}{T_i(\ell)}} + C_2 \frac{\log(\sqrt{N}\ell + 1)}{T_i(\ell)},$$

*with probability at least* $1 - \frac{7}{N\ell}$.

**Proof:** We first establish that the estimates $\hat{v}_{i,\ell},\ \ell \leq L$ are unbiased i.i.d estimates of the true parameter $v_i$ for all products. To prove this we show that the moment generating function of $\hat{v}_{i,\ell}$ conditioned on $S_\ell$ only depends on the parameter $v_i$ and not on the offered assortment $S_\ell$, there by establishing that estimates are independent and identically distributed. Using the moment generating function, we show that $\hat{v}_{i,\ell}$ is a geometric random variable with mean $v_i$, i.e., $E(\hat{v}_{i,\ell}) = v_i$. Using this observation and the classical multiplicative Chernoff-Hoeffding bounds to geometric random variables, we can arrive at Lemma 3.1

**Validity of the optimistic assortment.**

The next step in the regret analysis is to Lemma 3.1 to prove similar, though slightly weaker, properties for the estimates $\tilde{R}_\ell(S)$. Lemma 3.2 provides the precise statement.

**Lemma 3.2** *Suppose $S^* \in \mathcal{S}$ is the assortment with highest expected revenue, and Algorithm 2 offers $S_\ell \in \mathcal{S}$ in each epoch $\ell$. Then, for every epoch $\ell$, we have*

$$\tilde{R}_\ell(S_\ell) \geq \tilde{R}_\ell(S^*) \geq R(S^*, v) \textit{ with probability at least } 1 - \frac{6}{\ell}.$$

**Proof:** First, we show that estimated revenue is an upper confidence bound on the optimal revenue, i.e., $R(S^*, v)$ is bounded by $\tilde{R}_\ell(S_\ell)$ with high probability. The proof for these properties involves careful use of the structure of MNL model to show that the value of $\tilde{R}_\ell(S_\ell)$ is equal to the highest expected revenue achievable by any feasible assortment, among all instances of the problem with parameters in the range $[0, v_i^{\mathsf{UCB}}], i = 1, \ldots, n$. Since the actual parameters lie in this range with high probability, we have $\tilde{R}_\ell(S_\ell)$ is at least $R(S^*, v)$ with high probability.

**Bounding the regret.**

The final part of the analysis is to bound the regret in each epoch. First, we use the fact that $\tilde{R}_\ell(S_\ell)$ is an upper bound on $R(S^*, v)$ to bound the loss due to offering the assortment $S_\ell$. In particular, we show that the loss is bounded by the difference between the "optimistic" revenue estimate, $\tilde{R}_\ell(S_\ell)$, and the actual expected revenue, $R(S_\ell)$. Lemma 3.3 provides the precise statements of above properties.

**Lemma 3.3** *If $r_i \in [0, 1]$, there exists constants $C_1$ and $C_2$ such that for every $\ell = 1, \cdots, L$, we have*

$$(1 + \textstyle\sum_{j \in S_\ell} v_j)(\tilde{R}_\ell(S_\ell) - R(S_\ell, v)) \leq C_1 \sqrt{\frac{v_i \log(\sqrt{N}\ell + 1)}{|\mathcal{T}_i(\ell)|}} + C_2 \frac{\log(\sqrt{N}\ell + 1)}{|\mathcal{T}_i(\ell)|},$$

*with probability at least* $1 - \frac{13}{\ell}$.

**Proof:**The expected revenue function satisfies a certain kind of Lipschitz condition. Specifically, the difference between the estimated, $\tilde{R}_\ell(S_\ell)$, and expected revenues, $R_\ell(S_\ell)$, is bounded by the sum of errors in parameter estimates for the products, $|v_{i,\ell}^{\mathsf{UCB}} - v_i|$. This observation in conjunction with the "optimistic estimates" property will let us bound the regret as an aggregated difference between estimated revenues and expected revenues of the offered assortments. Noting that we have already computed convergence rates between the parameter estimates earlier, we can extend them to show that the estimated revenues converge to the optimal revenue.

## 3.5 Computational Study and Results

Here, we present a study that examines the robustness of Algorithm 2 with respect to the instance separability. We consider a parametric instance (see (28)), where the separation between the revenues of the optimal assortment and next best assortment is specified by the parameter $\epsilon$ and compare the performance of Algorithm 2 for different values of $\epsilon$.

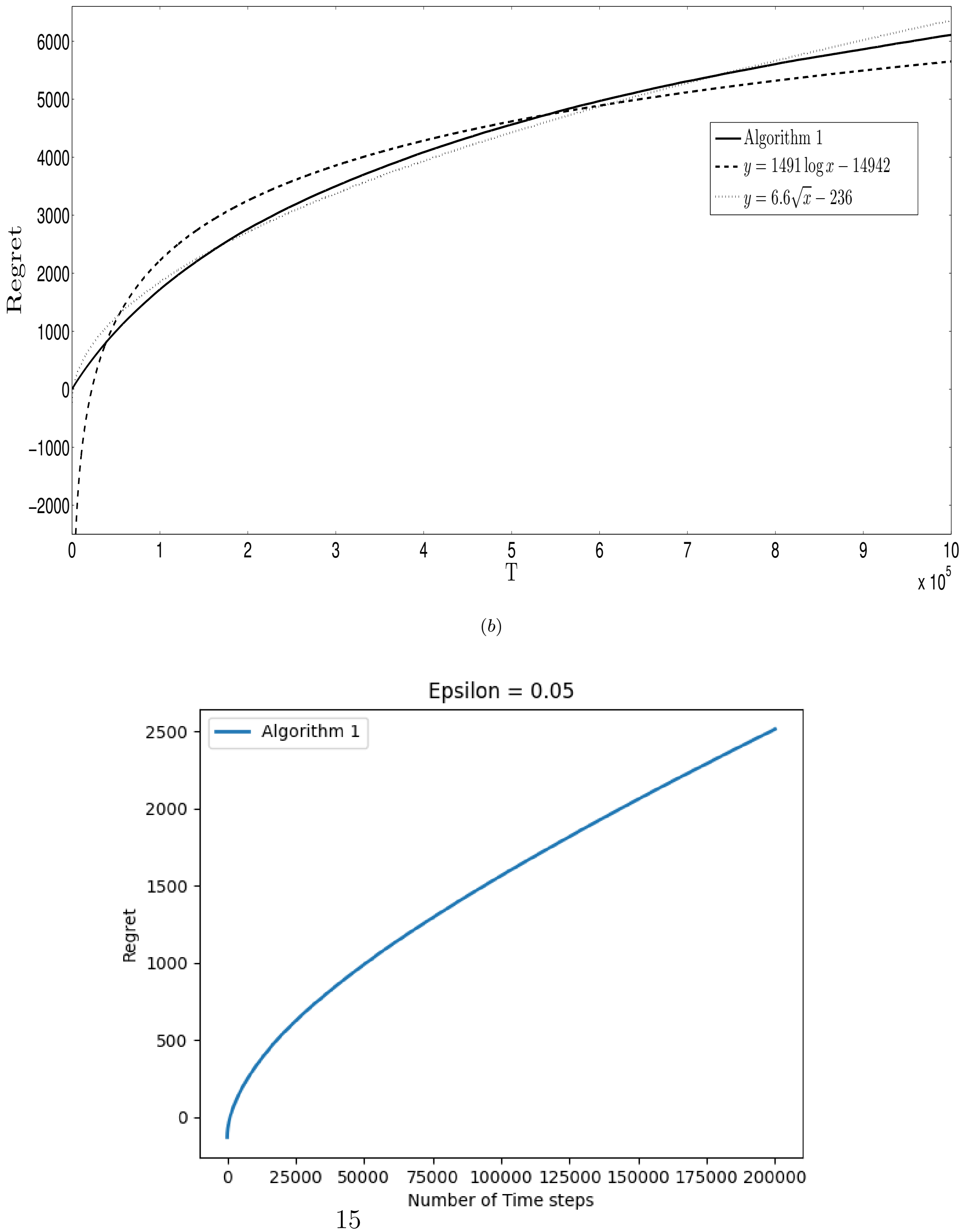
**Experimental setup 1.** We consider the parametric MNL setting with $N = 10$, $K = 4$, $r_i = 1$ for all $i$ and utility parameters $v_0 = 1$ and for $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else }, \end{cases} \tag{11}$$

where $0 < \epsilon < 0.25$, specifies the difference between revenues corresponding to the optimal assortment and the next best assortment. Note that this problem has a unique optimal assortment, $\{1, 2, 9, 10\}$ with an expected revenue of $1 + 4\epsilon/2 + 4\epsilon$ and next best assortment has revenue of $1 + 3\epsilon/2 + 3\epsilon$. We consider four different values for $\epsilon$, $\epsilon = \{0.05, 0.25\}$, where higher value of $\epsilon$ corresponds to larger separation, and hence an "easier" problem instance.

In the paper they ran 100 independent simulations of Algorithm 1 each for $10^6$ time steps. In my implementation I ran the 10 independent simulations each for $2 * 0^5$ time steps. From Figure 1 and 2, I was able to replicate and verify the results that the paper claims.

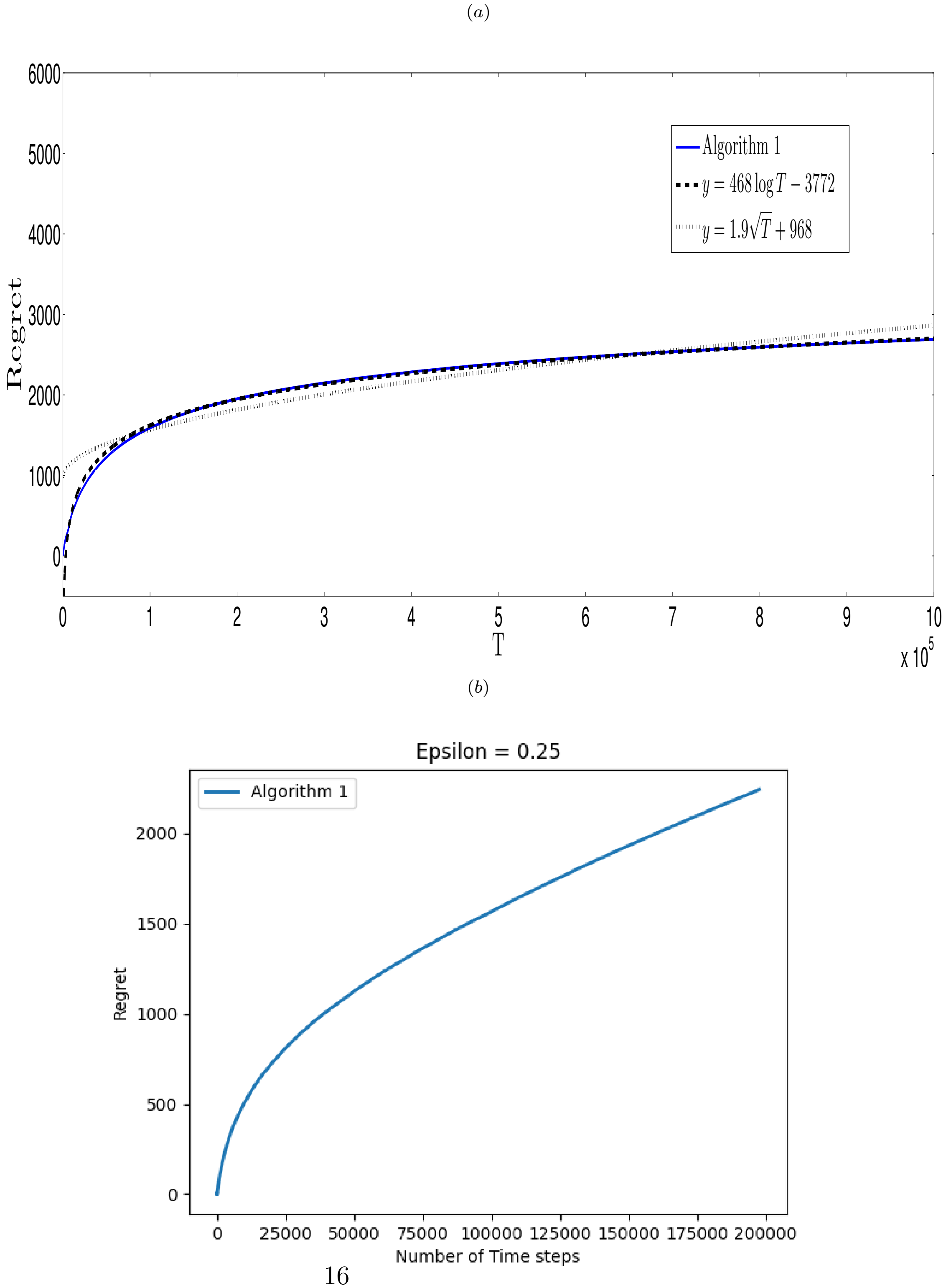
**Experimental Setup 2**: We consider the MNL setting with N = 2 and K = 1 and $v_1 = 0.1$ and $v_2 = 0.2$. A single run of the algorithm for $2 * 10^5$ time steps.



(a)

(b)

Figure 1: Best fit for the regret of Algorithm 2 on the parametric instance (28). The graphs (a), (b) illustrate the dependence of the regret on $T$ for "separation gap," $\epsilon = 0.05$ from the paper and my implementation respectively



$(a)$

$(b)$

Figure 2: Best fit for the regret of Algorithm 2 on the parametric instance (28). The graphs (a), (b) illustrate the dependence of the regret on $T$ for "separation gap," $\epsilon = 0.25$ from the paper and my implementation respectively

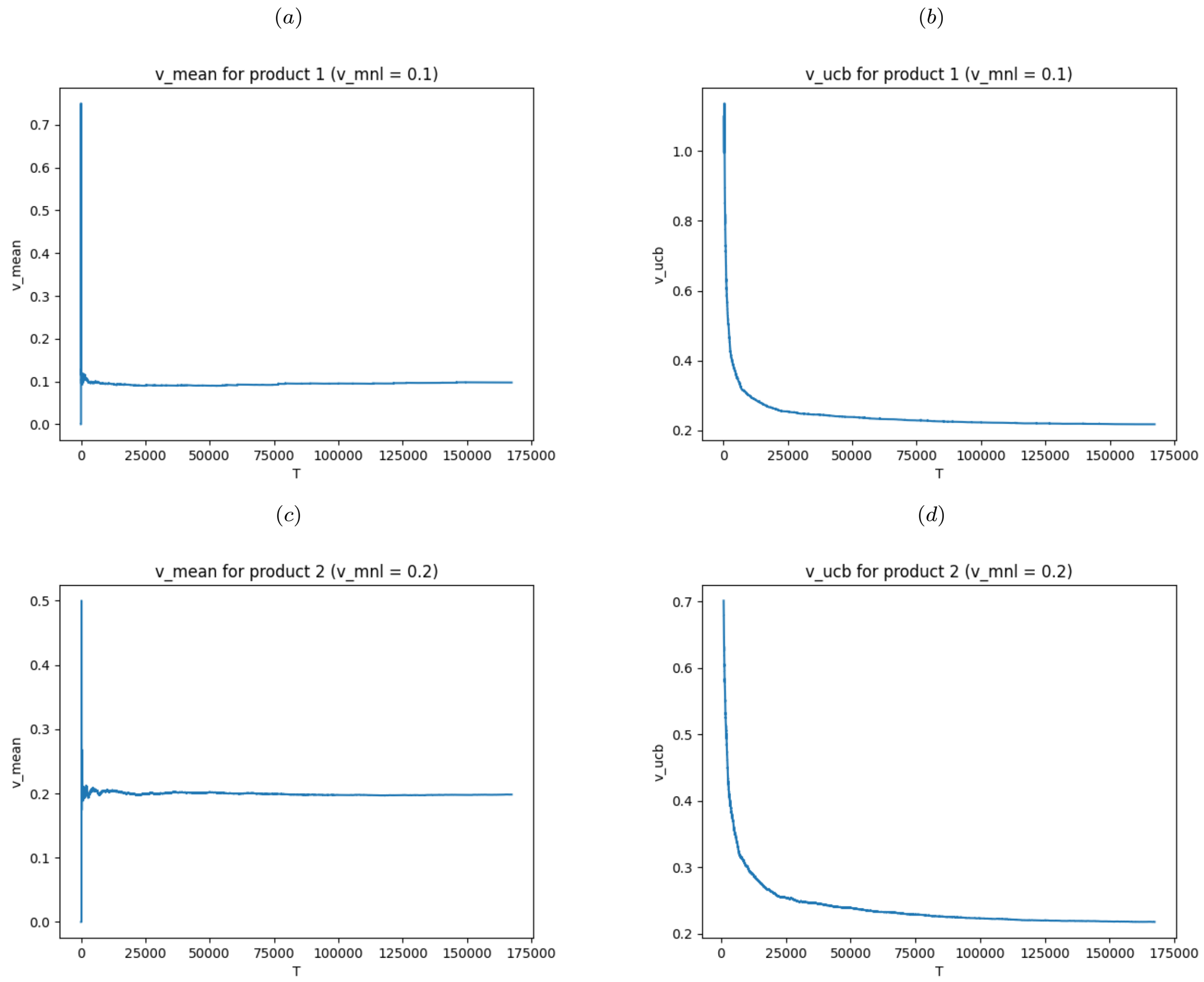


Figure 3: Algorithm 2 on the parametric instance of experiment setup 2. The graphs (a), (b), (c) and (d) illustrate how $v_{mean}$ and $v_{UCB}$ approach the actual $v_{mnl}$ parameters of 0.1 and 0.2

# 4 MDM Choice Models

## 4.1 Marginal Distribution Choice Models

A behavioral model for random utility is specified. In choice modeling, the semi-parametric approach assumes the modeler has some limited information about the distribution for the error terms, like the marginal distributions. Choice probabilities are estimated based on marginal distribution that meets predefined conditions, avoiding the need for complex integral calculations.
In MDM Models, utility of a consumer i for product j is defined as

$$U_{ij} = V_{ij} + \epsilon_{ij}. \tag{12}$$

The consumer will choose a product that will maximize his utility.

$$Z(U_i) = max \sum_{j \in N} U_{ij} y_{ij}, \sum_{j \in N} y_{ij} = 1, y_{ij} \in \{0, 1\} \forall j \in N (13)$$

The retailer will maximise the expected utility of the choices made by the consumer. The parameter $\theta^*$ would be:

$$\theta^* = argmax_{\theta \in \Theta} E(Z(U)) \tag{14}$$

## 4.2 Uniform Distribution

The cumulative probability for uniform distribution [0,1] is given by:

$$F_k(\alpha) = \begin{cases} 0, & \text{if } \alpha \leq 0 \\ \alpha, & \text{if } 0 < \alpha \leq 1 \\ 1, & \text{if } \alpha > 1 \end{cases} \tag{15}$$

Under MDM, the optimality conditions yield the choice probabilities of ith consumer and jth alternative in an assortment $S_l$, $P_{ij}$ would be,

$$P_{ij} = 1 - F_{ij}(\lambda_i - v_i j) \tag{16}$$

where the Lagrange multiplier $\lambda_i$ satisfies the following normalization condition:

$$\sum_{j \in S_l} P_{ij} = \sum_{j \in S_l} 1 - F_{ij}(\lambda_i - v_{ij}) \tag{17}$$

On solving for $\lambda$ in 17 we get,

$$\lambda_i = \frac{1 - K - \sum_{j \in S_l} v_{ij}}{K} \tag{18}$$

If we substitute $\lambda$ back in 16 we get.

$$P_{ij} = \frac{1 - \sum_{j \in S_l} v_{ij} + K v_{ij}}{K} \tag{19}$$

The corresponding revenue function for $P_{ij}$,

$$R(S_l, v) = \frac{\sum_{j \in S_l} r_j (1 + \sum_{j \in S_l} v_j + K v_{ij})}{K} \tag{20}$$

where $r_j$ is the revenue associated with the jth product.

## 4.3 Gamma Distribution

The cumulative probability for Gamma distribution with parameters $\alpha$ and $\lambda$ is given by:

$$F(x) = \int_0^{\lambda x} \frac{t^{\alpha-1} exp(-t) dt}{\Gamma(\alpha)} \tag{21}$$

Let $\alpha = 1$ and $\lambda = 1$ then, Under MDM, the optimality conditions yield the choice probabilities of ith consumer and jth alternative in an assortment $S_l$, $P_{ij}$ would be,

$$\begin{aligned} P_{ij} &= 1 - F_{ij}(\lambda_i - v_{ij}) \\ &= 1 - \int_0^{(\lambda_i - v_{ij})} e^{-t} dt \\ &= 1 - [-e^{-t}]_0^{(\lambda_i - v_{ij})} \\ &= e^{-(\lambda_i - v_{ij})} \end{aligned}$$

Therefore for Gamma distribution with $\lambda = 1$ and $\alpha = 1$

$$P_{ij} = e^{-(\lambda_i - v_{ij})} \tag{22}$$

where the Lagrange multiplier $\lambda_i$ satisfies the following normalization condition:

$$\sum_{j\in S_l} P_{ij} = 1$$
$$\sum_{j\in S_l} e^{-(\lambda_i - v_{ij})} = 1$$
$$e^{-\lambda_i}\sum_{j\in S_l} e^{v_{ij}} = 1$$

Therefore,

$$\lambda_i = ln(\sum_{j\in S_l} e^{v_{ij}}) \tag{23}$$

If we substitute $\lambda_i$ back in 22 we get,

$$P_{ij} = e^{-(ln(\sum_{j\in S_l} e^{v_{ij}}) - v_{ij})}. \tag{24}$$

The corresponding revenue function for $P_{ij}$,

$$R(S_l, v) = \sum_{j\in S_l} r_j e^{-[ln(\sum_{j\in S_l} e^{v_{ij}}) - v_{ij}]} \tag{25}$$

where $r_j$ is the revenue associated with the jth product.

## 4.4 MDM UCB Bandit Algorithm

In order to incorporate MDM choice model to the existing Algorithm 2, I modified the revenue function in the optimization step and the choice probability distribution in the choice generating function.
Uniform Distribution MDM,

$$\begin{aligned} S_\ell &= \underset{S\in\mathcal{S}}{\text{argmax}}\ \tilde{R}_\ell(S) \\ &= \underset{S\in\mathcal{S}}{\text{argmax}} \frac{\sum_{j\in S} r_j(1 + \sum_{j\in S} v^{UCB}_{ij,l-1} + |S| v^{UCB}_{ij,l-1})}{|S|} \end{aligned}$$

Gamma Distribution MDM,

$$\begin{aligned} S_\ell &= \underset{S\in\mathcal{S}}{\text{argmax}}\, \tilde{R}_\ell(S) \\ &= \underset{S\in\mathcal{S}}{\text{argmax}} \sum_{j\in S} r_j e^{-[ln(\sum_{j\in S} e^{v^{UCB}_{ij,l-1}}) - v^{UCB}_{ij,l-1}]} \end{aligned}$$

**Results**

- **Experimental setup 3.** We consider the parametric Uniform Distribution setting with $N = 10$, $K = 4$, $r_i = 1$ for all $i$ and utility parameters $v_0 = 1$ and for $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else }, \end{cases} \tag{26}$$

- The algorithm explored only 3 subsets even after $10^5$ epochs. This was probably because the UCB Bounds from Algorithm 1 need to be tuned to Uniform Distribution.

- **Experimental setup 4.** We consider the parametric Gamma Distribution setting with $N = 10$, $K = 4$, $r_i = 1$ for all $i$ and utility parameters $v_0 = 1$ and for $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else }, \end{cases} \tag{27}$$

- The algorithm explored only 1 subset even after $10^5$ epochs. This was probably because the UCB Bounds from Algorithm 1 need to be tuned to Gamma Distribution.

## 4.5 MDM $\epsilon$ greedy Bandit algorithm

An alternate approach could be a naive epsilon greedy approach as follows

**Algorithm 2** $\epsilon$ greedy approach to solve MDM Bandit problem

```
**Initialization:** $v_{i,0}^{\mathsf{UCB}} = 1$ for all $i = 1, \ldots, N$
$t = 1$ ; $\ell = 1$ keeps track of the time steps and total number of epochs respectively
**while** $t < T$ **do**
Generate a random probability $\epsilon$

  **if** $\epsilon < 0.4$ **then Exploit**
Compute $S_\ell = \underset{S \in \mathcal{S}}{\text{argmax}}\ \tilde{R}_\ell(S, \bar{v}_{i,\ell})$
Offer assortment $S_\ell$, observe the purchasing decision, $c_t$ of the consumer
      **if** $c_t = 0$ **then**
Compute $\hat{v}_{i,\ell} = \sum_{t \in E_\ell} 1(c_t = i)$, no. of consumers who preferred $i$ in epoch $\ell$, for all $i \in S_\ell$
Update $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until $\ell$ that offered product $i$
Update $\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in T_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
      **else**
$E_\ell = E_\ell \cup t$, time indices corresponding to epoch $\ell$
      **end if**
$t = t + 1$

  **else Explore**
Generate a random subset $S_\ell$ such that $|S_l| \leq K$
Offer assortment $S_\ell$, observe the purchasing decision, $c_t$ of the consumer
      **if** $c_t = 0$ **then**
Compute $\hat{v}_{i,\ell} = \sum_{t \in E_\ell} 1(c_t = i)$, no. of consumers who preferred $i$ in epoch $\ell$, for all $i \in S_\ell$
Update $\mathcal{T}_i(\ell) = \{\tau \leq \ell \,|\, i \in S_\ell\}, T_i(\ell) = |\mathcal{T}_i(\ell)|$, no. of epochs until $\ell$ that offered product $i$
Update $\bar{v}_{i,\ell} = \frac{1}{T_i(\ell)} \sum_{\tau \in T_i(\ell)} \hat{v}_{i,\tau}$, sample mean of the estimates
      **else**
$E_\ell = E_\ell \cup t$, time indices corresponding to epoch $\ell$
      **end if**
$t = t + 1$
  **end if**
**end while**=0
```

**Results**

- **Experimental setup 5.** We consider the parametric Uniform Distribution setting with $N = 10$, $K = 4$, $r_i = 1$ for all $i$ and utility parameters $v_0 = 1$ and for $i = 1, \ldots, N$,

$$v_i = \begin{cases} 0.25 + \epsilon, & \text{if } i \in \{1, 2, 9, 10\} \\ 0.25, & \text{else }, \end{cases} \tag{28}$$

- The number of epochs it takes for the values of $\bar{v}_{i,\ell}$ to converge to actual parameter value is way too high and the implementation was taking up a lot of time to run.

# 5 Conclusion and Future Work

- The paper presents a novel algorithm for dynamic assortment selection under the multinomial logit (MNL) model, which balances exploration and exploitation without prior knowledge of problem parameters. The algorithm is computationally efficient and nearly optimal in performance, adapting to the complexity of the problem instance.

- The primary focus was on universally applicable algorithms, considering only the setting where each product has its own utility parameter. Extensions to factor models or heterogeneous consumer settings, and the design of Thompson Sampling-based approaches for the MNL-Bandit problem, are suggested as important future research directions. The paper also acknowledges the challenge of analyzing Thompson Sampling-based algorithms due to their combinatorial nature.

- Exploring factor models that describe products via a small number of features, designing algorithms for heterogeneous consumers, and developing Thompson Sampling-based approaches for the MNL-Bandit problem are identified as potential areas for future research.

- The algorithm's adaptability to problem complexity and its performance without requiring prior knowledge of parameters have significant practical implications for dynamic assortment optimization in retail and online advertising.

- An open is to derive the UCB for $v_i^{UCB}$ for Uniform and Gamma distribution and plug that value into Algorithm 1 UCB updates and the regret bounds for MDM with the new $v_i^{UCB}$ need to analaysed mathematically.

# 6 References

Agrawal, Avadhanula, Goyal and Zeevi: MNL-Bandit: A Dynamic Learning Approach to Assortment Selection Submitted to Operations Research 00(0),https://arxiv.org/abs/1706.03880

Mishra, V. K., Natarajan, K., Padmanabhan, D., Teo, C.-P., Li, X. (2014). On Theoretical and Empirical Aspects of Marginal Distribution Choice Models. Management Science, 60(6), 1511–1531. http://www.jstor.org/stable/42919617

Natarajan, Karthik Song, Miao Teo, Chung. (2009). Persistency Model and Its Applications in Choice Modeling. Management Science. 55. 453-469. 10.1287/mnsc.1080.0951.

Jain, Shweta Bhat, Satyanath Ghalme, Ganesh Padmanabhan, Divya Narahari, Yadati. (2016). Mechanisms with learning for stochastic multi-armed bandit problems. Indian Journal of Pure and Applied Mathematics. 47. 229-272. 10.1007/s13226-016-0186-3.