MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

Thompson Sampling for the MNL-Bandit

Shipra Agrawal
Industrial Engineering and Operations Research, Columbia University, New York, NY. sa3305@columbia.edu

Vashist Avadhanula
Decision, Risk and Operations, Columbia Business School, New York, NY. vavadhanula18@gsb.columbia.edu

Vineet Goyal
Industrial Engineering and Operations Research, Columbia University, New York, NY. vg2277@columbia.edu

Assaf Zeevi
Decision, Risk and Operations, Columbia Business School, New York, NY. assaf@gsb.columbia.edu

We consider a dynamic combinatorial optimization problem, where at each time step the decision maker selects a subset of cardinality K from N possible items, and observes a feedback in the form of the index of one of the items in said subset, or none. Each of the N items is ascribed a certain value (reward), which is collected if the item is chosen. This problem is motivated by that of assortment selection in online retail, where items are products. Akin to that literature, it is assumed that the choice of the item given the subset is governed by a Multinomial Logit (MNL) choice model whose parameters are a priori unknown. The objective of the decision maker is to maximize the expected cumulative rewards over a finite horizon T, or alternatively, minimize the regret relative to an oracle that knows the MNL choice model parameters. We formulate this problem as a multi-armed bandit problem that we refer to as the MNL-Bandit problem. We present a Thompson Sampling based algorithm for this problem and show that it achieves near-optimal regret as well as attractive empirical performance.

Key words: Thompson Sampling, Exploration-Exploitation, Multinomial Logit Choice Model
MSC2000 subject classification:
OR/MS subject classification: Primary: ; secondary:
History: Abstract of the preliminary version of this work has been published in the proceedings of Conference on Learning Theory (COLT) 2017. Pre-print of this document is available on Arxiv at `https://arxiv.org/pdf/1706.00977.pdf`. The contents of this document are also included in the dissertation of Vashist Avadhanula, which is available at `https://academiccommons.columbia.edu/doi/10.7916/d8-6c67-1n16`.

1. Introduction.

1.1. Overview of the problem. In the canonical stochastic multi-armed Bandit (MAB) problem, at each time step $t = 1, \ldots, T$ a single arm is chosen out of the set $\{1, \ldots, N\}$, in response to which a noisy reward, characteristic of that arm, is observed. The objective is to minimize the gap between the performance of a policy and that of an oracle that selects the arm with the highest expected reward in each round; this gap is often referred to as the regret. In many instances of the problem, probably good performance can be achieved by employing a design principle known as "optimism in the face of uncertainty." A prime example is the widely studied family of upper confidence bound policies (UCB), see, e.g., Auer et al. [6], that suitably balance the exploration-exploitation tension inherent in MAB problems.

In this paper, we consider a combinatorial variant of this problem where at each time step $t = 1, \cdots, T$ the player selects a subset of cardinality K from the index set of N arms, after which s/he either observes the reward associated with one of the arms in this subset, along with the identity of the arm that generated the reward or observes no reward at all. One can think of the "no reward" option as feedback that manifests from a "null arm" that is augmented to each subset.

In our set up, the rewards are deterministic and taken to be the problem primitives, but the identity of the arm within the chosen subset that yields the reward (or the "null" arm that yields no reward) is drawn from a probability distribution on the index set of cardinality $K+1$ (the K arms plus the "null" arm). In this paper the distribution is specified by means of a multinomial logit model (MNL) whose parameters are not known to the player a priori and can only be inferred over time via the revealed indices (and rewards); the term MNL-Bandit refers to these salient features. The objective, as in the traditional MAB formulation, is to develop playing strategies that try to come close to the performance of an oracle that solves the full information offline problem, or in other words, minimize the regret.

The problem as stated above is of central importance in a variety of application domains, notable examples include display-based online advertising (where one can only display K ads on a content page, selected from a typically large universe N of feasible choices), and more generally the design of product recommendation engines (where K limits the number of products that can be displayed to the consumer in response to a search query). While its mathematical formulation is not new, there is surprisingly little antecedent literature on this problem; the review below will expound a bit on its history and related strands of work.

1.2. Main contributions. Relying on structural properties of the MNL model, we develop the first computationally efficient Thompson Sampling (TS) approach to the MNL-Bandit and study its theoretical properties. TS belongs to a Bayesian class of learning algorithms where a (repeatedly updated) posterior distribution governs the sampling of actions in each stage; some further context and relevant work is discussed in the literature review. Its main attractive feature vis-a-vis UCB-type policies, is improved (numerical) regret performance that stems essentially from more efficient exploration (UCB tends to be conservative and over-explore). However, for the MNL-bandit problem, TS turns out to be far more difficult to analyze theoretically, and presents significant computational challenges that hinder efficient implementation. These stem primarily from the computational demands involved with the calculation of posterior distribution, and the "closed-loop" structure that links observations, updates, and actions.

The main challenges associated with the analysis of TS algorithms are overcome through several key components. First, a carefully chosen prior distribution on the parameters of the MNL model is put in place, and this allows for efficient and tractable posterior updating under the MNL-bandit feedback. A second key ingredient in our approach is a two-moment approximation of the posterior which is embedded within a normal family. A fundamental consequence of this embedding is the ability to correlate samples which plays a central role in the performance of our algorithm. The methods developed in this paper highlight these key attributes, and present a blueprint to address these issues that we hope will be more broadly applicable and form the basis for further work in the intersection of combinatorial optimization and machine learning.

Our main theoretical contribution is a worst-case (prior-free) regret bound on the performance of our proposed algorithm which exhibits an order of $O(\sqrt{NT}\log TK)$; the bound is non-asymptotic, the "big oh" notation is used for brevity and simplicity. This regret bound is independent of the parameters of the MNL choice model and hence holds uniformly over all problem instances of size N, K. Moreover, it is essentially best possible due to a lower bound of $\Omega(\sqrt{NT})$ established recently by Wang et al. [34] for the MNL-Bandit. Hence our TS algorithm achieves regret-optimal performance up to logarithmic terms. We note in passing that this bound is comparable to the existing upper bound of $\tilde{O}(\sqrt{NT})$ obtained in Agrawal et al. [2] for a UCB-based algorithm for the MNL-Bandit. However, as will be seen in the sequel, numerical results demonstrate that our TS-based approach significantly outperforms the UCB-based approach of Agrawal et al. [2] and the results in this paper provide the first theoretical analysis for a TS based approach.

Organization. The remainder of the paper is organized as follows. The current section concludes with a brief review of related literature that helps place our contributions in context. We provide the mathematical formulation of our problem in Section 2. In Section 3, we present our Thompson Sampling algorithm for the MNL-Bandit, and in Section 4, we prove the main result, namely, that our algorithm achieves an $O(\sqrt{NT}\log TK)$ regret upper bound. Section 5 demonstrates the empirical efficiency of our algorithm design.

1.3. Related work A basic pillar in the MNL-Bandit problem is the MNL choice model, originally introduced (independently) by Luce [17] and Plackett [25]; see also Train [33], McFadden [19], Ben-Akiva and Lerman [8] for further discussion and survey of other commonly used choice models. The MNL model is the most widely used choice model for capturing substitution effects that are a significant element in our problem. Rusmevichientong et al. [26] and Sauré and Zeevi [31] were the first two papers we are aware of to consider a dynamic learning problem for the MNL-Bandit in the context of a retail assortment optimization problem. Both papers are predicated on an "explore first and exploit later" approach and their algorithms are parameter-dependent. Specifically, assuming knowledge of a "gap" value (between the optimal and the next-best subsets), Sauré and Zeevi [31] establish an asymptotic $O(N\log T)$ regret bound. (This assumption is akin to the "separated arm" case in the MAB setting.) In a more recent paper, Agrawal et al. [2] develop a UCB-like algorithm which does not rely on the a priori knowledge of this gap and show that this algorithm achieves a worst-case regret bound of $O(\sqrt{NT\log T})$. A regret lower bound of $\Omega(\sqrt{NT/K})$ for this problem is also presented in this work, which was subsequently improved to $\Omega(\sqrt{NT})$ in a recent work by Wang et al. [34] establishing the near optimality of this UCB algorithm.

Subsequent follow up works, Chen et al. [9, 10, 11], Cheung and Simchi-Levi [12], Saha and Gopalan [30], Feng et al. [14], Miao and Chao [20, 21] and Oh and Iyengar [22, 23] consider different variants of the MNL-Bandit problem. The works of Chen et al. [10], Miao and Chao [21] and Oh and Iyengar [22] considers the more general contextual variant of the MNL-Bandit problem. These papers builds on Agrawal et al. [2] to develop UCB based approaches and establish worst-case regret bounds of $\tilde{O}(d\sqrt{T})$, where d is the dimension of contexts. However, the algorithms and regret bounds presented in these papers are dependent on certain problem parameters. Following an initial conference version of this paper, the works of Cheung and Simchi-Levi [12], Miao and Chao [20] and Oh and Iyengar [23] developed Thompon Sampling based approaches for contextual variations of the MNL-bandit problem. These works achieve a Bayesian regret bound of $\tilde{O}(d\sqrt{T})$ that are dependent on problem parameters. In this work, we provide worst-case regret bounds that are independent of the problem parameters and hold uniformly overall problem instances of size N, K. Feng et al. [14] and Saha and Gopalan [30] considers the best arm identification variant of the MNL-Bandit problem, where the focus is only on exploration to identify the best K items. In this work, we focus on optimally balancing the exploration-exploitation tradeoffs, a completely different setting from the aforementioned works. Chen et al. [9] considers the variant of the MNL-Bandit where feedback from a small fraction of users is not consistent with the MNL choice model. The paper presents a near-optimal algorithm with a worst-case regret bound of $\tilde{O}(\epsilon K^2 T + \sqrt{NKT})$, where ϵ is the fraction of users for whom the feedback is corrupted. However, the algorithm developed by Chen et al. [9] which focuses on robustness to corruption is sub-optimal in the setting considered in this work ($\epsilon = 0$).

The basic idea of Thompson Sampling for MAB was introduced by Thompson [32]. TS starts with a prior on the underlying parameter space and as each arm is pulled and observations are collected, a posterior is updated and further actions are determined using draws from this posterior. Several recent studies (Oliver and Li [24], Graepel et al. [16], May et al. [18]) have demonstrated that TS significantly outperforms state of the art learning algorithms in practice, and over the past few years, TS has received renewed interest both in theoretical studies as well as a plethora of implementations. At the same time, TS based algorithms are notoriously difficult to analyze and theoretical work on TS is limited. To the best of our knowledge, Agrawal and Goyal [3] is the first work on TS in a traditional MAB setting that provides a finite-time worst-case regret bound independent of problem parameters; see also work by Russo and Van Roy [28] for Bayesian regret bounds.

A naive translation of the MNL-Bandit problem to the basic MAB problem setting would create $\binom{N}{K}$ "arms" (one for each offer set of size K). Managing such an exponentially large arm space is prohibitive for obvious reasons. Popular extensions of MAB for "large scale" problems include the linear bandit (e.g., Auer [5], Rusmevichientong and Tsitsiklis [27]) for which Agrawal and Goyal [4] present a TS-based algorithm and provide finite time regret bounds. However, these approaches do not apply directly to our problem, since the revenue corresponding to each chosen subset is not linear in problem parameters. Gopalan et al.

[15] consider a variant of MAB where one can play a subset of arms in each round and the expected reward is a function of rewards of the arms played. This setting is similar to the MNL-Bandit, though the regret bounds they develop are dependent on the instance parameters as well as the number of possible actions, which can be large in our combinatorial problem setting. Russo et al. [29] presents efficient heuristics to approximate the TS algorithm considered in Russo and Van Roy [28], however, it is not immediately clear if these approximate TS-based approaches facilitate theoretical analysis.

2. Problem Formulation. To formally state our problem, consider an option space containing N distinct elements, indexed by $1,\dots,N$ and their values, denoted by $r_1,\dots,r_N$, with r being the mnemonic for reward, though we will also use the term *revenue* in this context. We append the option space by an additional element indexed by "0", in order to represent the alternative available to the user of not selecting any of the options presented. We assume that for any offer set, $S \subset \{1,\dots,N\}$, the user selects only one of the offered alternatives or item 0, according to a Multinomial Logit (MNL) choice model. Under this model, given the offer set S, the probability that the user chooses item $i \in S$ is given by,

$$p_i(S) = \begin{cases} \dfrac{v_i}{v_0 + \sum_{j\in S} v_j}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $v_0,\dots,v_N$ are parameters of the MNL model. Without loss of generality, we can assume that $v_0 = 1$.

The expected reward or revenue corresponding to the offer set S, $R(S,\mathbf{v})$ is given by

$$R(S,\mathbf{v}) := \sum_{i\in S} r_i p_i(S) = \sum_{i\in S} \frac{r_i v_i}{1+\sum_{j\in S} v_j}. \tag{2}$$

We consider a setting where the decision maker can select at most K products in the offer set.

$$S^* := \max\{\, R(S,\mathbf{v}) | |S| \le K\}. \tag{3}$$

Such a cardinality constraint arises naturally in many applications. For instance, a publisher/retailer is constrained by the space for advertisements/products and has to limit the number that can be displayed simultaneously.

We can now formulate the MNL-bandit problem as follows. The problem proceeds in discrete sequential rounds $t = 1,\dots,T$ for some predetermined time horizon T. In each round t, the decision maker offers a K-cardinality subset of items $S_t \subseteq \{1,\dots,N\}$ and observes the user's choice $c_t \in S \cup \{0\}$. The probability distribution of c_t is given by the MNL choice model as described in (1). The MNL choice model parameters $v_1,\dots,v_N$ are apriori fixed but unknown to the decision maker. The objective is to design an algorithm that selects a (non-anticipating) sequence of offer sets in a path-dependent manner (namely, based on past choices and observed responses) to maximize cumulative expected reward over said horizon or, alternatively, minimize the *regret* defined as

$$\text{Reg}(T,\mathbf{v}) = \mathbb{E}\left[\sum_{t=1}^{T} R(S^*,\mathbf{v}) - R(S_t,\mathbf{v})\right], \tag{4}$$

where $R(S,\mathbf{v})$ is the expected reward when the offer set is S, and is as defined in (2). Here we make explicit the dependence of regret on the time horizon T and the parameter vector $\mathbf{v}$ of the MNL model, that determines the user preferences and choices.

3. Algorithm. In this section, we describe our posterior sampling (aka Thompson Sampling) based algorithm for the MNL-Bandit problem. The basic structure of Thompson Sampling involves maintaining a posterior on the unknown problem parameters, which is updated every time new feedback is obtained. At the beginning of every round, a sample set of parameters is generated from the current posterior distribution, and the algorithm selects the best offer set according to these sample parameters. In the MNL-Bandit problem, there is one unknown parameter v_i associated with each item. To adapt the TS algorithm for this problem, we would need to maintain a joint posterior for $(v_1, \ldots, v_N)$. However, updating such a joint posterior is non-trivial since the feedback observed in every round is a choice sample from the multinomial distribution. This depends on the subset S offered in that round. In particular, even if we initialize with an independent prior from a popular analytical family such as multivariate Gaussian, the posterior distribution after observing the MNL choice feedback will have a complex description. As a first step in addressing this challenge, we attempt to design a Thompson Sampling algorithm with independent priors. In particular, we leverage a sampling technique introduced in Agrawal et al. [2] that allows us to decouple individual parameters from the MNL choice feedback and provide unbiased estimates of these parameters. We can utilize these unbiased estimates to efficiently maintain independent conjugate Beta priors for the parameters v_i for each i. We present the details in Algorithm 1 below.

3.1. A TS algorithm with independent conjugate Beta priors Here we present the first version of our Thompson sampling algorithm, which will serve as an important building block for our main algorithm in Section 3.2. In this version, we maintain a Beta posterior distribution for each item $i = 1, \ldots, N$, which is updated as we observe users' choice of items from the offered subsets. A key challenge here is to choose priors that can be efficiently updated on observing user choice feedback, to obtain increasingly accurate estimates of parameters $\{v_i\}$. To address this, we use the sampling technique introduced in Agrawal et al. [2] to decouple estimates of individual parameters from the complex MNL feedback. The idea is to offer a set S multiple times; in particular, a chosen set S is offered repeatedly until the "outside option" is picked (in the online advertising application discussed earlier, this corresponds to displaying the same subset of ads repeatedly until we observe a user who does not click on any of the displayed ads). Proceeding in this manner, due to the structure of the MNL model, the average number of times an item i is selected provides an unbiased estimate of parameter v_i. Moreover, as derived in Agrawal et al. [2], the number of times an item i is selected is also independent of the displayed set and is a geometric distribution with success probability $1/(1+v_i)$ and mean v_i. This observation is used as the basis for our epoch based algorithmic structure and our choice of prior/posterior, as a conjugate to this geometric distribution.

Epoch based offerings: Our algorithm proceeds in epochs $\ell = 1, 2, \ldots$. An epoch is a group of consecutive time steps, where a set S_ℓ is offered repeatedly until the outside option is picked in response to offering S_ℓ. The set S_ℓ to be offered in epoch ℓ is picked at the beginning of the epoch based on the sampled parameters from the current posterior distribution; the construction of these posteriors and choice of S_ℓ is described in the next paragraph. We denote the group of time steps in an epoch as $\mathcal{E}_\ell$, which includes the time step at which an outside option was preferred.

The following lemmas provide important building blocks for our construction. Their proofs have been deferred to the appendix.

Lemma 1 (Agrawal et al. [2]) *Let $\tilde{v}_{i,\ell}$ be the number of times an item $i \in S_\ell$ is picked when the set S_ℓ is offered repeatedly until the outside option is picked. Then, $\tilde{v}_{i,\ell}$ forall ℓ, i are i.i.d geometric random variables with success probability $\frac{1}{1+v_i}$, and expected value v_i.*

Lemma 2 (Conjugate Priors) *For any $\alpha > 3, \beta > 0$ and $Y_{\alpha,\beta} \sim \mathsf{Beta}(\alpha, \beta)$, let $X_{\alpha,\beta} = \frac{1}{Y_{\alpha,\beta}-1}$ and $f_{\alpha,\beta}$ denote the probability distribution of random variable $X_{\alpha,\beta}$. If the prior distribution of v_i is $f_{\alpha,\beta}$, then after observing $\tilde{v}_{i,\ell}$, a geometric random variable with success probability $\frac{1}{v_i+1}$, the posterior distribution of v_i is given by,*

$$\mathbb{P}\left(v_i \middle| \tilde{v}_{i,\ell} = m\right) = f_{\alpha+1,\beta+m}(v_i).$$

Construction of conjugate prior/posterior: From Lemma 1, we have that for any epoch ℓ and for any item $i \in S_\ell$, the estimate $\tilde{v}_{i,\ell}$, the number of picks of item i in epoch ℓ is geometrically distributed with success probability $1/(1+v_i)$. Therefore, if we use the distribution of $1/\mathsf{Beta}(1,1) - 1$ as the initial prior for v_i, and then, in the beginning of epoch ℓ, from Lemma 2 we have that the posterior is distributed as $\frac{1}{\mathsf{Beta}(n_i(\ell),V_i(\ell))} - 1$, with $n_i(\ell)$ being the number of epochs the item i has been offered before epoch ℓ (as part of an assortment), and $V_i(\ell)$ being the number of times it was picked by the user.

Selection of subset to be offered: To choose the subset to be offered in epoch ℓ, the algorithm samples a set of parameters $\mu_1(\ell), \ldots, \mu_N(\ell)$ independently from the current posteriors and finds the set that maximizes the expected revenue as per the sampled parameters. In particular, the set S_ℓ to be offered in epoch ℓ is chosen as:

$$S_\ell := \underset{|S| \le K}{\operatorname{argmax}} R(S, \boldsymbol{\mu}(\ell)), \tag{5}$$

where the reward function $R(.,.)$ is given in (2). There are efficient polynomial time algorithms available to solve this optimization problem (e.g., Davis et al. [13], Avadhanula et al. [7] and Rusmevichientong et al. [26]).

The details of the above procedure are provided in Algorithm 1.

Algorithm 1 A TS algorithm for MNL-Bandit with Independent Beta priors

```
Initialization: For each item $i = 1, \cdots, N$, $V_i = 1$, $n_i = 1$.
$t = 1$, keeps track of the time steps
$\ell = 1$, keeps count of total number of epochs
while $t \le T$ do
    (a) (Posterior Sampling) For each item $i = 1, \cdots, N$, sample $\theta_i(\ell)$ from the $\mathsf{Beta}(n_i, V_i)$ and compute
        $\mu_i(\ell) = \frac{1}{\theta_i(\ell)} - 1$
    (b) (Subset Selection) Compute $S_\ell = \underset{|S| \le K}{\operatorname{argmax}} \, R(S, \boldsymbol{\mu}(\ell)) = \frac{\sum_{i \in S} r_i \mu_i(\ell)}{1 + \sum_{j \in S} \mu_j(\ell)}$
    (c) (Epoch-based offering)
        repeat
            Offer the set $S_\ell$, and observe the user choice $c_t$;
            Update $\mathcal{E}_\ell = \mathcal{E}_\ell \cup t$, time indices corresponding to epoch $\ell$; $t = t + 1$
        until $c_t = 0$
    (d) (Posterior update)
            For each item $i \in S_\ell$, compute $\tilde{v}_{i,\ell} = \sum_{t \in \mathcal{E}_\ell} \mathbb{I}(c_t = i)$, number of picks of item $i$ in epoch $\ell$.
            Update $V_i$ = $V_i + \tilde{v}_{i,\ell}$, $n_i = n_i + 1$, $\ell = \ell + 1$.
end while
```

Algorithm 1 presents some unique challenges for theoretical analysis. A worst-case regret analysis of Thompson Sampling-based algorithms for MAB typically relies on showing that the best-arm is optimistic at least once every few steps, in the sense that the parameter sampled from the posterior is better than the true parameter. Due to the combinatorial nature of our problem, such a proof approach requires showing that every few steps, all the K items in the optimal offer set have sampled parameters that are better than their true counterparts. However, Algorithm 1 samples the posterior distribution for each parameter *independently* in each round. This makes the probability of being optimistic exponentially small in K. In Section 3.2, we modify Algorithm 1 to address these challenges and in a manner amenable to theoretical analysis.

3.2. A TS algorithm with posterior approximation and correlated sampling

We address the challenge associated with the combinatorial nature of the MNL-Bandit by employing *correlated sampling* across items. To implement correlated sampling, we find it useful to approximate the Beta posterior distribution by a Gaussian distribution with approximately the same mean and variance as the former; what was referred to in the introduction as a two-moment approximation. This allows us to generate correlated samples from the N Gaussian distributions as linear transforms of a single standard Gaussian random variable. Under such correlated sampling, we can guarantee that the probability of all K optimal items are simultaneously optimistic is constant, as opposed to being exponentially small (in K) in the case of independent sampling. However, such correlated sampling reduces the overall variance of the maximum of N samples severely, thus inhibiting exploration. We "boost" the variance by taking K samples instead of a single sample of the standard Gaussian. The resulting variant of Thompson Sampling, therefore has three main modifications, posterior approximation through a Gaussian distribution; correlated sampling; and taking multiple samples (for "variance boosting"). We elaborate on each of these changes below.

Posterior approximation: First, we present the following result that helps us in approximating the posterior. Proof of the result has been deferred to the appendix.

Lemma 3 (Moments of the Posterior Distribution) *If X is a random variable distributed as* $\mathsf{Beta}(\alpha,\beta)$*, then*

$$\mathbb{E}\left(\tfrac{1}{X}-1\right)=\tfrac{\beta}{\alpha-1}, \quad \textit{and}\ \mathsf{Var}\left(\tfrac{1}{X}-1\right) \quad =\tfrac{\frac{\beta}{\alpha-1}\left(\frac{\beta}{\alpha-1}+1\right)}{\alpha-2}.$$

We approximate the posterior distributions used in Algorithm 1 for each MNL parameter v_i, by a Gaussian distribution with approximately the same mean and variance given in Lemma 3. In particular, let

$$\hat{v}_i(\ell):=\frac{V_i(\ell)}{n_i(\ell)},\ \ \hat{\sigma}_i(\ell):=\sqrt{\frac{50\hat{v}_i(\ell)(\hat{v}_i(\ell)+1)}{n_i(\ell)}}+75\frac{\sqrt{\log TK}}{n_i(\ell)},\quad \ell=1,2,\ldots \tag{6}$$

where $n_i(\ell)$ is the number of epochs item i has been offered before epoch ℓ, and $V_i(\ell)$ being the number of times it was picked by the user. We will use $\mathcal{N}\left(\hat{v}_i(\ell),\hat{\sigma}_i^2(\ell)\right)$ as the posterior distribution for item i in the beginning of epoch ℓ. The Gaussian approximation of the posterior facilitates efficient correlated sampling from posteriors that plays a key role in avoiding the theoretical challenges in analyzing Algorithm 1.

Correlated sampling: Given the posterior approximation by Gaussian distributions, we correlate the samples by using a common standard normal variable and constructing our posterior samples as an appropriate transform of this common standard normal. More specifically, in the beginning of an epoch ℓ, we generate a sample from the standard normal distribution, $\theta\sim\mathcal{N}(0,1)$ and the posterior sample for item i, is generated as $\hat{v}_i(\ell)+\theta\hat{\sigma}_i(\ell)$. Intuitively, this allows us to generate sample parameters for $i=1,\ldots,N$ that are either simultaneously large or simultaneously small, thereby, boosting the probability that the sample parameters for *all* the K items in the best offered set are optimistic (i.e., the sampled parameter values are higher than the true parameter values).

Multiple (K) samples: The correlated sampling decreases the joint variance of the sample set. More specifically, if θ_i were sampled independently from the standard normal distribution for every i, then for any epoch ℓ, we have that

$$\mathsf{Var}\left(\max_{i=1,\cdots,N}\{\hat{v}_i(\ell)+\theta\hat{\sigma}_i(\ell)\}\right)\leq\mathsf{Var}\left(\max_{i=1,\cdots,N}\{\hat{v}_i(\ell)+\theta_i\hat{\sigma}_i(\ell)\}\right).$$

In order to boost this joint variance and ensure sufficient exploration, we modify the procedure to generate multiple sets of samples. In particular, in the beginning of an epoch ℓ, we now generate K independent samples from the standard normal distribution, $\theta^{(j)}\sim\mathcal{N}(0,1), j=1,\ldots,K$. And then for each j, a sample parameter set is generated as:

$$\mu_i^{(j)}(\ell):=\hat{v}_i(\ell)+\theta^{(j)}\hat{\sigma}_i(\ell),\quad i=1,\ldots,N,$$

Then, we use the largest valued samples

$$\mu_i(\ell) := \max_{j=1,\cdots,K} \ \mu_i^{(j)}(\ell), \forall i,$$

to decide the assortment to offer in epoch ℓ,

$$S_\ell := \arg\max_{S\in\mathcal{S}} \{R(S, \boldsymbol{\mu}(\ell))\}$$

We describe the details formally in Algorithm 2.

Algorithm 2 A TS algorithm for MNL-Bandit with Gaussian approximation and correlated sampling

```
Input parameters: $\alpha = 50$, $\beta = 75$
Initialization: $t = 0$, $\ell = 0$, $n_i = 0$ for all $i = 1, \cdots, N$.
for each item, $i = 1, \cdots, N$ do
    Offer item $i$ to users until the user selects the "outside option". Let $\tilde{v}_{i,1}$ be the number of times item $i$ was offered. Update: $V_i = \tilde{v}_{i,1} - 1$, $t = t + \tilde{v}_{i,1}$, $\ell = \ell + 1$ and $n_i = n_i + 1$.
end for
while $t \leq T$ do
    (a) (Correlated Sampling) for $j = 1, \cdots, K$
            Sample $\theta^{(j)}(\ell)$ from the distribution $\mathcal{N}(0,1)$ and let $\theta_{\max}(\ell) = \max_{j=1,\cdots,K} \theta^{(j)}(\ell)$;  update $\hat{v}_i = \frac{V_i}{n_i}$.
            For each item $i \leq N$, compute $\mu_i^{(j)}(\ell) = \hat{v}_i + \theta_{\max}(\ell) \cdot \left(\sqrt{\frac{\alpha \hat{v}_i(\hat{v}_i+1)}{n_i}} + \frac{\beta\sqrt{\log TK}}{n_i}\right)$.
        end
    (b) (Subset selection) Same as step (b) of Algorithm 1.
    (c) (Epoch-based offering) Same as step (c) of Algorithm 1.
    (d) (Posterior update) Same as step (d) of Algorithm 1.
end while
```

Intuitively, the second-moment approximation provided by Gaussian distribution and the multiple samples taken in Algorithm 2 may make the posterior converge slowly and increase exploration. However, the correlated sampling may compensate for these effects by reducing the variance of the maximum of N samples, and therefore reducing the overall exploration. In Section 5, we illustrate some of these insights through numerical simulations. Here, correlated sampling is observed to provide significant improvements as compared to independent sampling, and while posterior approximation by Gaussian distribution has little impact.

4. Regret Analysis

We prove an upper bound on the regret of Algorithm 2 for the MNL-Bandit problem, under the following assumption.

Assumption 1 For every item $i \in \{1, \ldots, N\}$, the MNL parameter v_i satisfies $v_i \leq v_0 = 1$.

This assumption is equivalent to the outside option being more preferable compared to any other item. This assumption holds for many applications including display advertising, where users do not click on any of the displayed ads more often than not. Our main theoretical result is the following upper bound on the regret of Algorithm 2.

Theorem 1 *For any instance $\mathbf{v} = (v_0, \cdots, v_N)$ of the MNL-Bandit problem with N products, $r_i \in [0,1]$, and satisfying Assumption 1, the regret of Algorithm 2 in time T is bounded as,*

$$\mathit{Reg}(T, \mathbf{v}) \le C_1 \sqrt{NT} \log TK + C_2 N \log^2 TK,$$

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

4.1. Proof Outline. In this section, we provide a proof sketch for Theorem 1. We break down the expression for total regret

$$\text{Reg}(T, \mathbf{v}) := \mathbb{E}\left[\sum_{t=1}^{T} R(S^*, \mathbf{v}) - R(S_t, \mathbf{v})\right],$$

into regret per epoch, and rewrite it as follows:

$$\text{Reg}(T, \mathbf{v}) = \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} |\mathcal{E}_\ell| \left(R(S^*, \mathbf{v}) - R(S_\ell, \boldsymbol{\mu}(\ell))\right)\right]}_{\text{Reg}_1(T,\mathbf{v})} + \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} |\mathcal{E}_\ell| \left(R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v})\right)\right]}_{\text{Reg}_2(T,\mathbf{v})}, \tag{7}$$

where $|\mathcal{E}_\ell|$ is the number of periods in epoch ℓ, and S_ℓ is the set repeatedly offered by our algorithm in epoch ℓ. We bound the two terms: $\text{Reg}_1(T, \mathbf{v})$ and $\text{Reg}_2(T, \mathbf{v})$ separately.

Since S_ℓ is chosen as the optimal set for the MNL instance with parameters $\boldsymbol{\mu}(\ell)$, the first term $\text{Reg}_1(T, \mathbf{v})$ is essentially the difference between the optimal revenue of the true instance and the optimal revenue of the sampled instance. This term contributes no regret if the revenues corresponding to the sampled instances are optimistic, i.e., if $R(S_\ell, \boldsymbol{\mu}(\ell)) \ge R(S^*, \mathbf{v})$. Unlike optimism under uncertainty approaches such as UCB, this property is not directly ensured by our Thompson Sampling-based algorithm. To bound this term, we utilize the anti-concentration properties of the posterior, as well as the dependence between samples for different items. In particular, we use these properties to prove that at least one of the K sampled instances is optimistic "often enough."

The second term $\text{Reg}_2(T, \mathbf{v})$ captures the difference in reward from the offered set S_ℓ when evaluated on sampled parameters in comparison to the true parameters. We bound this by utilizing the concentration properties of the posterior distributions. It involves showing that for the sets that are played often, the posterior will converge quickly so that revenue on the sampled parameters will be close to that on the true parameters.

In what follows, we will first highlight three key results involved in proving Theorem 1. In Section 4.2 we will combine these properties and follow the above outline to prove Theorem 1.

Structural properties of the optimal revenue. The first step in our regret analysis is to leverage the structure of the MNL model to establish two key properties of the optimal expected revenue. These properties project the non-linear reward function of the MNL choice into its parameter space and help us focus on analyzing the posterior distribution of the parameters. In the first property, which we refer to as restricted monotonicity, we note that the optimal expected revenue is monotone in the MNL parameters. The second property, is a Lipschitz property of the expected revenue function that bounds the difference between the expected revenue corresponding to two different MNL parameters in terms of the difference in individual parameters. Lemma 4 provides the precise statement.

Lemma 4 (Properties of the Optimal Revenue) *For any $\mathbf{v} \in \mathcal{R}_+^n$, let S^* be an optimal assortment for MNL instance with parameters $\mathbf{v}$, i.e. $S^* = \underset{S:|S| \le K}{\mathit{argmax}}\, R(S, \mathbf{v})$. Then, for any $\mathbf{w} \in \mathcal{R}_+^n$, we have:*

1. *(Restricted Monotonicity) If $v_i \le w_i$ for all $i = 1, \cdots, N$. Then, $R(S^*, \mathbf{w}) \ge R(S^*, \mathbf{v})$.*

2. *(Lipschitz Property)* $|R(S^*, \mathbf{v}) - R(S^*, \mathbf{w})| \leq \dfrac{\sum_{i\in S^*} |v_i - w_i|}{1 + \sum_{j\in S^*} v_j}$.

Proof. We will first prove the restricted monotonicity property and extend the analysis to prove the Lipschitz property.

Restricted Monotonicity. We prove the result by first showing that for any $j \in S^*$, we have $R(S^*, \mathbf{w}^j) \geq R(S^*, \mathbf{v})$, where $\mathbf{w}^j$ is vector $\mathbf{v}$ with the j^{th} component increased to w_j, i.e. $w^j_i = v_i$ for all $i \neq j$ and $w^j_j = w_j$. We can use this result iteratively to argue that increasing each parameter of MNL to the largest possible value increases the value of $R(S, \mathbf{w})$.

If there exists $j \in S$ such that $r_j < R(S)$, then removing the product j from assortment S yields higher expected revenue contradicting the optimality of S. Therefore, we have

$$r_j \geq R(S) \text{ forall } j \in S.$$

Multiplying by $(v_j - w_j)(\sum_{i\in S/j} w_i + 1)$ on both sides of the above inequality and re-arranging terms, we can show that $R(S^*, \mathbf{w}^j) \geq R(S^*, \mathbf{v})$.

Lipschitz. Following the above analysis, we define sets $\mathcal{I}(S^*)$ and $\mathcal{D}(S^*)$ as

$$\begin{aligned} \mathcal{I}(S^*) &= \{i | i \in S^* \text{ and } v_i \geq w_i\} \\ \mathcal{D}(S^*) &= \{i | i \in S^* \text{ and } v_i < w_i\}, \end{aligned}$$

and vector $\mathbf{u}$ as,

$$u_i = \begin{cases} w_i & \text{if } i \in \mathcal{D}(S^*), \\ v_i & \text{otherwise.} \end{cases}$$

By construction of $\mathbf{u}$, we have $u_i \geq v_i$ and $u_i \geq w_i$ for all i. Therefore from the restricted monotonicity property, we have

$$\begin{aligned} R(S^*, \mathbf{v}) - R(S^*, \mathbf{w}) &\leq R(S^*, \mathbf{u}) - R(S^*, \mathbf{w}) \\ &\leq \frac{\sum\limits_{i\in S^*} r_i u_i}{1 + \sum\limits_{j\in S^*} u_j} - \frac{\sum\limits_{i\in S^*} r_i w_i}{1 + \sum\limits_{j\in S^*} u_j}, \\ &\leq \frac{\sum\limits_{i\in S^*} (u_i - w_i)}{1 + \sum\limits_{j\in S^*} u_j} \leq \frac{\sum_{i\in S^*} |v_i - w_i|}{1 + \sum_{j\in S^*} v_j}. \end{aligned}$$

The Lipschitz property in Lemma 4 follows from the definition of u_i. This completes the proof. □

Concentration of the posterior distribution. The next step in the regret analysis is to show that as more observations are made, the posterior distributions concentrate around their means, which in turn concentrate around the true parameters. More specifically, we have the following two results.

Lemma 5 (Bounds on Gaussian Distribution) *For any $\ell \leq T$ and $i \in \{1, \cdots, N\}$, we have for any $r > 0$,*

$$\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right) \leq \frac{1}{r^4 K^3},$$

where $\hat{\sigma}_i(\ell)$ is as defined in (6).

Lemma 6 (Multiplicative Chernoff Bound) *If $v_i \leq 1$ for all $i = 1, \cdots, N$, then for any $m, \rho > 0$, $\ell \in \{1, 2, \cdots\}$ and $i \in \{1, \cdots, N\}$ we have,*

1. $\mathbb{P}\left(|\hat{v}_i(\ell) - v_i| > 4\sqrt{\frac{\hat{v}_i(\ell)(\hat{v}_i(\ell)+1)m\log(\rho+1)}{n_i(\ell)}} + \frac{24m\log(\rho+1)}{n_i(\ell)}\right) \leq \frac{5}{\rho^m}.$
2. $\mathbb{P}\left(|\hat{v}_i(\ell) - v_i| \geq \sqrt{\frac{12v_i m\log(\rho+1)}{n_i(\ell)}} + \frac{24m\log(\rho+1)}{n_i(\ell)}\right) \leq \frac{4}{\rho^m}.$

The above results indicate that for any item i and at the beginning of any epoch ℓ, the difference between the sample from the posterior distribution $\mu_i(\ell)$ and the true parameter v_i is bounded and is decreasing over time. Lemma 5 follows from the properties of the Gaussian distribution and Lemma 6 is an adaptation of Chernoff bounds for our setting. For the sake of continuity, we defer the proof of these concentration results to Appendix A.1. Leveraging the Lipschitz property of the optimal revenue, this concentration of sample parameters around their true values will help us prove that the difference between the expected revenue of the offer set S_ℓ corresponding to the sampled parameters, $\boldsymbol{\mu}(\ell)$, and the true parameters, $\mathbf{v}$ also becomes smaller with time. In particular, we have the following result.

Lemma 7 (Towards Reg$_2(T,\mathbf{v})$ Bound) *For any epoch ℓ, if $S_\ell = \underset{S:|S|\leq K}{argmax}\, R(S,\boldsymbol{\mu}(\ell))$*

$$\mathbb{E}\left\{(1+\sum_{j\in S_\ell} v_j)\left[R(S_\ell,\boldsymbol{\mu}(\ell)) - R(S_\ell,\mathbf{v})\right]\right\} \leq \mathbb{E}\left[C_1\sum_{i\in S_\ell}\sqrt{\frac{v_i\log TK}{n_i(\ell)}} + C_2\frac{\log TK}{n_i(\ell)}\right],$$

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

The concentration property of the posterior distribution allows us to bound the second term, $\mathsf{Reg}_2(T,\mathbf{v})$ in (7). Therefore to bound the regret, it suffices to bound the difference between the optimal revenue $R(S^*,\mathbf{v})$ and the expected revenue of the offer set corresponding to sampled parameters $R(S_\ell,\boldsymbol{\mu}(\ell))$.

Anti-Concentration of the posterior distribution. We refer to an *epoch ℓ as optimistic* if the expected revenue of the optimal set corresponding to the sampled parameters is higher than the expected revenue of the optimal set corresponding to true parameters, i.e., $R(S^*,\boldsymbol{\mu}(\ell)) \geq R(S^*,\mathbf{v})$. Any epoch that is not optimistic is referred to as a *non-optimistic epoch.* Since S_ℓ is an optimal set for the sampled parameters, we have $R(S_\ell,\boldsymbol{\mu}(\ell)) \geq R(S^*,\boldsymbol{\mu}(\ell))$. Hence, for any optimistic epoch ℓ, the difference between the expected revenue of the offer set corresponding to sampled parameters $R(S_\ell,\boldsymbol{\mu}(\ell))$ and the optimal revenue $R(S^*,\mathbf{v})$ is bounded by zero. This suggests that as the number of optimistic epochs increases, the term $\mathsf{Reg}_1(T,\mathbf{v})$ decreases.

The central technical component of our analysis is showing that the regret over non-optimistic epochs is "small." More specifically, we prove that there are only a "small" number of non-optimistic epochs. From the restricted monotonicity property of the optimal revenue (see Lemma 4), we have that an epoch ℓ is optimistic if every sampled parameter, $\mu_i(\ell)$ is at least as high as the true parameter v_i for every item i in the optimal set S^*. Recall that each posterior sample $\mu_i^{(j)}(\ell)$, is generated from a Gaussian distribution, whose mean concentrates around the true parameter v_i. We can use this observation to conclude that any sampled parameter will be greater than the true parameter with constant probability, i.e. $\mu_i^{(j)}(\ell) \geq v_i$. However, to show that an epoch is optimistic, we need to show that sampled parameters for *all* the items in S^* are larger than the true parameters. This is where the correlated sampling feature of our algorithm plays a key role. We use the dependence structure between samples for different items in the optimal set, and variance boosting (by a factor of K) to prove an upper bound of roughly $1/K$ on the number of consecutive epochs between two optimistic epochs. More specifically, we have the following result.

Lemma 8 (Spacing of optimistic epochs) *Let $\mathcal{E}^{\mathsf{An}}(\tau)$ denotes the set of consecutive epochs between an optimistic epoch τ and the subsequent optimistic epoch τ'. For any $p \in [1,2]$, we have,*

$$\mathbb{E}^{1/p}\left[\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p\right] \leq \frac{e^{12}}{K} + 30^{1/p}.$$

Proof. Note that for any non-negative discrete random variable, X, we have $E(X) = \sum_x P(X \geq x)$. Hence, we will first establish a lower bound on the probability $\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p \geq q\right\}$ and use the preceding fact to obtain a bound on the moments of the number of non-optimistic epochs.

For the sake of brevity, let $r = \lfloor q^{1/p} \rfloor$ and $z = \sqrt{\log(rK+1)}$. Hence, we have,

$$\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right|^p \geq q\right\} = \mathbb{P}\left\{\left|\mathcal{E}(\tau)\right| \geq r\right\}.$$

By definition, $[\mathcal{E}^{\mathsf{An}}(\tau) < r]$ implies that one of the epochs $\tau+1, \cdots, \tau+r$ is optimistic. More specifically we have,

$$\begin{aligned}\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| > r\right\} &= 1 - \mathbb{P}\left(\left\{\ \{\mu_i(\ell) \geq v_i \text{ for all } i \in S^*\} \text{ for some } \ell \in (\tau, \tau+r]\right\}\right),\\ &\leq 1 - \mathbb{P}\left(\left\{\ \{\mu_i(\ell) \geq \hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \geq v_i \text{ for all } i \in S^*\} \text{ for some } \ell \in (\tau, \tau+r]\right\}\right).\end{aligned}$$

For the sake of brevity, let A_ℓ denote the event that the sampled parameter for every item in the optimal set is larger than z standard deviations away from the mean of the posterior distribution. Furthermore, let B_ℓ denote the event that the true parameter of every item in the optimal set is smaller than mean of the posterior distribution plus z times the standard deviation of the posterior distribution. More specifically we have,

$$\begin{aligned}A_\ell &= \{\mu_i(\ell) \geq \hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \ \text{ for all } i \in S^*\},\\ B_\ell &= \{\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) \geq v_i \ \text{ for all } i \in S^*\}.\end{aligned}$$

Therefore we have,

$$\begin{aligned}\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| \geq r\right\} &\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c \cup B_\ell^c\right),\\ &\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) + \sum_{\ell=\tau+1}^{\tau+r} \mathbb{P}(B_\ell^c), \qquad (8)\\ &\leq \mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) + \sum_{i \in S^*} \mathbb{P}\left(\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) < v_i\right).\end{aligned}$$

where the last two inequalities follow from the union bound. Note that from the concentration property of the posterior distribution (see Lemma 6), the probability of every event in the above inequality is small. In particular, substituting $m = 3.1$ and $\rho = rK$ in Lemma 6 and using the fact that $rK \leq TK$ we obtain,

$$\mathbb{P}\left(\hat{v}_i(\ell) + z\hat{\sigma}_i(\ell) < v_i\right) \leq \frac{1}{(rK)^{3.1}}. \qquad (9)$$

We will now use the tail bounds for Gaussian random variables to bound the probability $\mathbb{P}(A_\ell^c)$. For any Gaussian random variable, Z with mean μ and standard deviation σ, we have,

$$Pr(Z > \mu + x\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} e^{-x^2/2}.$$

Note that by design of Algorithm 2, $\mu_i(\ell) = \hat{v}_i(\ell) + \hat{\sigma}_i(\ell) \max\limits_{j \leq K} \theta^{(j)}(\ell)$, where $\theta^{(j)}(\ell)$ are i.i.d standard normal random variables. Therefore, we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{\ell=\tau+1}^{\tau+r} A_\ell^c\right) &= \mathbb{P}\Big(\theta^{(j)}(\ell) \leq z \ \text{ for all } \ell \in (\tau, \tau+r] \text{ and for all } j = 1, \cdots, K\Big),\\ &\overset{(a)}{\leq} \left[1 - \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\log rK}}{\log rK + 1} \cdot \frac{1}{\sqrt{rK}}\right)\right]^{rK}, \qquad (10)\\ &\overset{(b)}{\leq} \exp\left(-\frac{r^{1/2}}{\sqrt{2\pi}} \frac{2\sqrt{\log rK}}{4\log rK + 1}\right),\\ &\overset{(c)}{\leq} \frac{1}{(rK)^{2.2}} \ \text{ for any } r \geq \frac{e^{12}}{K},\end{aligned}$$

where inequality (a) follows from the tail bounds, inequality (b) follows from the fact that $1-x \leq e^{-x}$ for all $x \geq 0$ and inequality (c) follows from the fact that $\exp\left(-\sqrt{x/2\pi \log x}\right) \leq 1/x^{2.2}$ for any $x \geq e^{12}$.

Hence from (8), (9), and (10) we have ,

$$\mathbb{P}\left\{\left|\mathcal{E}^{\mathsf{An}}(\tau)\right| \geq r\right\} \leq \frac{1}{(rK)^{2.1}} + \frac{1}{(rK)^{2.2}} \text{ for any } r \geq \frac{e^{12}}{K}.$$

The result follows from the above inequality, definition of r and the fact that $\sum_{x=1}^{\infty} \frac{1}{x^y}$ is constant for any $y > 1$. This completes the proof. □.

4.2. Putting it all together: Proof of Theorem 1 In this section, we will utilize the above properties and follow the outline discussed in Section 4.1 to complete the proof of Theorem 1. For the sake of brevity we will use the following notation for the rest of this section.

- For any offer set S, $V(S) := \sum_{i \in S} v_i$
- For any $\ell, \tau \leq L$, define ΔR_ℓ and $\Delta R_{\ell,\tau}$ in the following manner

$$\begin{aligned}\Delta R_\ell &:= (1+V(S_\ell))\left[R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \mathbf{v})\right]\\ \Delta R_{\ell,\tau} &:= (1+V(S_\tau))\left[R(S_\ell, \boldsymbol{\mu}(\ell)) - R(S_\ell, \boldsymbol{\mu}(\tau))\right]\end{aligned}$$

- Let $\mathcal{A}_0$ denote the complete set Ω and for all $\ell = 1, \ldots, L$, define events $\mathcal{A}_\ell$ as

$$\mathcal{A}_\ell = \left\{|\hat{v}_i(\ell) - v_i| \geq \sqrt{\frac{24 v_i \log(\ell+1)}{n_i(\ell)}} + \frac{48 \log(\ell+1)}{n_i(\ell)} \text{ for some } i = 1, \cdots, N\right\}$$

We bound the regret by bounding both the terms in (7). We first focus on bounding the second term, $\mathsf{Reg}_2(T, \mathbf{v})$ and then extend the analysis to bound, $\mathsf{Reg}_1(T, \mathbf{v})$.

Bounding $\mathsf{Reg}_2(T, \mathbf{v})$: Note that conditioned on S_ℓ, the length of the ℓ^{th} epoch, $|\mathcal{E}^{\mathsf{Al}}|$ is a geometric random variable with probability of success $p_0(S_\ell) = 1/(1+V(S_\ell))$. Therefore using conditional expectations, we can reformulate $\mathsf{Reg}_2(T, \mathbf{v})$ as,

$$\mathsf{Reg}_2(T, \mathbf{v}) = \mathbb{E}\left\{\sum_{\ell=1}^{L} \Delta R_\ell\right\}. \tag{11}$$

Noting that $\mathcal{A}_\ell$ is a "low probability" event, we analyze the regret in two scenarios: one on $\mathcal{A}_\ell$, and another on $\mathcal{A}_\ell^c$. More specifically, for any ℓ

$$\begin{aligned}\mathbb{E}\left(\Delta R_\ell\right) &= \mathbb{E}\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right],\\ &\leq \frac{K+1}{\ell^2} + \mathbb{E}\left[\Delta R_\ell \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right],\end{aligned} \tag{12}$$

where the last inequality follows from Lemma 6 and the fact that both $R(S_\ell, \boldsymbol{\mu}(\ell))$ and $R(S_\ell, \mathbf{v})$ are bounded by 1 and $V(S_\ell) \leq K$. Therefore from Lemma 4 it follows that,

$$\begin{aligned}\mathbb{E}\left[\Delta R_\ell \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right] &\leq \mathbb{E}\left[\sum_{i \in S_\ell} |\mu_i(\ell) - v_i| \cdot \mathbb{1}(\mathcal{A}_{\ell-1}^c)\right].\\ &\leq \mathbb{E}\left[\sum_{i \in S_\ell} |\mu_i(\ell) - \hat{v}_i(\ell)|\right] + \mathbb{E}\left[\sqrt{\frac{24 v_i \log(\ell+1)}{n_i(\ell)}} + \frac{48 \log(\ell+1)}{n_i(\ell)}\right],\end{aligned} \tag{13}$$

where the last inequality follows from the definition of event $\mathcal{A}_\ell$ and triangle inequality.

Using Lemma 5 we show that the first term in above inequality, which is difference between the sampled parameter and the mean of the posterior distribution is bounded (see Corollary 6). Therefore, from (11), (12), (13), Corollary A.1 and Lemma 6, we have,

$$Reg_2(T,\mathbf{v}) \le C_1\mathbb{E}\left(\sum_{\ell=1}^{L}\sum_{i\in S_\ell}\sqrt{\frac{v_i\log TK}{n_i(\ell)}}\right) + C_2\mathbb{E}\left(\sum_{\ell=1}^{L}\sum_{i\in S_\ell}\frac{\log TK}{n_i(\ell)}\right), \tag{14}$$

where C_1 and C_2 are absolute constants. If T_i denote the total number of epochs product i is offered, then we have,

$$\begin{aligned} Reg_2(T,\mathbf{v}) &\overset{(a)}{\le} C_2 N\log^2 TK + C_1\mathbb{E}\left(\sum_{i=1}^{n}\sqrt{v_iT_i\log TK}\right), \\ &\overset{(b)}{\le} C_2 N\log^2 TK + C_1\sum_{i=1}^{N}\sqrt{v_i\log{(TK)}\mathbb{E}(T_i)}. \end{aligned} \tag{15}$$

Inequality (a) follows from the observation that $L \le T$, $T_i \le T$, $\displaystyle\sum_{n_i(\ell)=1}^{T_i}\frac{1}{\sqrt{n_i(\ell)}} \le \sqrt{T_i}$ and $\displaystyle\sum_{n_i(\ell)=1}^{T_i}\frac{1}{n_i(\ell)} \le \log T_i$, while Inequality (b) follows from Jensen's inequality.

Since that expected epoch length conditioned on the event $S = S_\ell$ is $1+V(S_\ell)$, we have, $\sum v_i\mathbb{E}(T_i) \le T$. To obtain the worst case upper bound, we maximize the bound in Equation (15) subject to the above condition. Therefore, we have

$$Reg_2(T,\mathbf{v}) \le C_1\sqrt{NT\log TK} + C_2 N\log^2 TK). \tag{16}$$

We now focus on the first term in (7), $\mathsf{Reg}_1(T,\mathbf{v})$.

Bounding $\mathsf{Reg}_1(T,\mathbf{v})$**:** Let $\mathcal{T}$ denote the set of optimistic epochs. Recall that $\mathcal{E}^{\mathsf{An}}(\ell)$ is the set of non-optimistic epochs between ℓ^{th} epoch and the subsequent optimistic epoch. Therefore, we can reformulate $Reg_1(T,\mathbf{v})$ as,

$$\mathsf{Reg}_1(T,\mathbf{v}) = \mathbb{E}[\sum_{\ell=1}^{L}\mathbb{1}(\ell\in\mathcal{T})\cdot\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}|\mathcal{E}_\tau|(R(S^*,\mathbf{v}) - R(S_\tau,\boldsymbol{\mu}(\tau)))]$$

Note that for any ℓ, by design S_ℓ is the optimal set for the sampled parameters, i.e., $R(S_\ell,\boldsymbol{\mu}(\ell)) \ge R(S^*,\boldsymbol{\mu}(\ell))$. From the restricted monotonicity property, for any $\ell\in\mathcal{T}$, we have $R(S^*,\boldsymbol{\mu}(\ell)) \ge R(S^*,\mathbf{v})$. Therefore, it follows that,

$$\begin{aligned} \mathsf{Reg}_1(T,\mathbf{v}) \le\ & \mathbb{E}\left[\sum_{\ell=1}^{L}\mathbb{1}(\ell\in\mathcal{T})\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}|\mathcal{E}_\tau|(R(S_\ell,\boldsymbol{\mu}(\ell)) - R(S_\tau,\boldsymbol{\mu}(\tau)))\right], \\ \overset{(a)}{\le}\ & \mathbb{E}\left[\sum_{\ell=1}^{L}\mathbb{1}(\ell\in\mathcal{T})\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}|\mathcal{E}_\tau|(R(S_\ell,\boldsymbol{\mu}(\ell)) - R(S_\ell,\boldsymbol{\mu}(\tau)))\right], \\ \overset{(b)}{\le}\ & \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}\Delta R_{\ell,\tau}\right] \end{aligned} \tag{17}$$

where inequality (a) follows from the fact S_τ is the optimal assortment for the sampled parameters $\boldsymbol{\mu}(\tau)$ and inequality (b) follows from the observation that the expected length of the τ^{th} epoch conditioned on event $S = S_\tau$ is $1+V(S_\tau)$. Following the approach of bounding $\mathsf{Reg}_2(T,\mathbf{v})$, we analyze the first term, $\mathsf{Reg}_1(T,\mathbf{v})$ in two scenarios, one on $\mathcal{A}_\ell$ and one on $\mathcal{A}^c_\ell$. More specifically,

$$\begin{aligned}
\mathbb{E}\left(\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)} \Delta R_{\ell,\tau}\right) &= \mathbb{E}\left[\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)} \Delta R_{\ell,\tau}\mathbb{1}(\mathcal{A}_{\ell-1}) + \Delta R_{\ell,\tau}\mathbb{1}(\mathcal{A}^c_{\ell-1})\right], \\
&\overset{(a)}{\leq} (K+1)\mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)|\mathbb{1}(\mathcal{A}_{\ell-1})] + \Delta R_{\ell,\tau}\mathbb{1}(\mathcal{A}^c_{\ell-1})], \\
&\overset{(b)}{\leq} (K+1)\mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)|\mathbb{1}(\mathcal{A}_{\ell-1})] + \mathbb{E}[\mathbb{1}(\mathcal{A}^c_{\ell-1})\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}\sum_{i\in S_\ell}|\mu_i(\ell)-\mu_i(\tau)|], \\
&\overset{(c)}{\leq} (K+1)\mathbb{E}[|\mathcal{E}^{\mathsf{An}}(\ell)|\mathbb{1}(\mathcal{A}_{\ell-1})] + \mathbb{E}\left[\mathbb{1}(\mathcal{A}^c_{\ell-1})\sum_{\tau\in\mathcal{E}^{\mathsf{An}}(\ell)}\sum_{i\in S_\ell}|\mu_i(\ell)-v_i| + |\mu_i(\tau)-v_i|\right],
\end{aligned} \tag{18}$$

where, inequality (a) follows from the fact that $R(S_\ell,\boldsymbol{\mu}(\ell))$ and $R(S_\ell,\boldsymbol{\mu}(\tau))$ are bounded by 1 and $V(S_\tau)\leq K$; inequality (b) follows from Lemma 4; and inequality (c) follows from the triangle inequality.

Following the approach of Bounding $\mathsf{Reg}_2(T,\mathbf{v})$, specifically along the lines of (13) and Corollary A.1, we can show that

$$\mathbb{1}(\mathcal{A}^c_{\ell-1})|\mu_i(\ell)-v_i| \leq C_1\sqrt{\frac{v_i\log TK}{n_i(\ell)}} + \frac{\log TK}{n_i(\ell)}.$$

Since $\tau\geq\ell$ we have $n_i(\ell)\leq n_i(\tau)$. Therefore, from (17), (18) and Lemma 6 we obtain the following inequality.

$$\mathsf{Reg}_1(T,\mathbf{v}) \leq \mathbb{E}\left[\sum_{\ell\in\mathcal{T}}|\mathcal{E}^{\mathsf{An}}(\ell)|\sum_{i\in S_\ell}\left(C_1\sqrt{\frac{v_i\log TK}{n_i(\ell)}} + C_2\frac{\log TK}{n_i(\ell)}\right)\right], \tag{19}$$

for some absolute constants C_1 and C_2. If $|\mathcal{E}^{\mathsf{An}}(.)|$ is constant, then bounding the above inequality is similar to bounding $\mathsf{Reg}_1(T,\mathbf{v})$ (see (14)). In the remainder of this section, we will show how to utilize Lemma 8 to bound $\mathsf{Reg}_1(T,\mathbf{v})$. From Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{\ell\in\mathcal{T}}\sum_{i\in S_\ell}|\mathcal{E}^{\mathsf{An}}(\ell)|C_1\sqrt{\frac{v_i\log TK}{n_i(\ell)}}\right] &\leq C_1\sum_{\ell}\sum_{i\in S_\ell}\mathbb{E}^{1/2}\left[|\mathcal{E}^{\mathsf{An}}(\ell)|^2\right]\cdot\mathbb{E}^{1/2}\left[\frac{v_i\log TK}{n_i(\ell)}\right], \\
\mathbb{E}\left[\sum_{\ell\in\mathcal{T}}\sum_{i\in S_\ell}|\mathcal{E}^{\mathsf{An}}(\ell)|C_2\frac{\log TK}{n_i(\ell)}\right] &\leq C_2\sum_{\ell}\sum_{i\in S_\ell}\mathbb{E}^{1/2}\left[|\mathcal{E}^{\mathsf{An}}(\ell)|^2\right]\mathbb{E}^{1/2}\left[\frac{\log^2 TK}{n_i^2(\ell)}\right].
\end{aligned}$$

Therefore from Lemma 8 for some absolute constant C, we have,

$$\begin{aligned}
\mathsf{Reg}_1(T,\mathbf{v}) &\leq \frac{C}{K}\left(\sum_{\ell}\sum_{i\in S_\ell}\mathbb{E}^{1/2}\left[\frac{v_i\log TK}{n_i(\ell)}\right] + \sum_{\ell}\sum_{i\in S_\ell}\mathbb{E}^{1/2}\left[\frac{\log^2 TK}{n_i^2(\ell)}\right]\right), \\
&\leq \frac{C}{K}\left(\sqrt{TK\mathbb{E}\left[\sum_{\ell}\sum_{i\in S_\ell}\frac{v_i\log TK}{n_i(\ell)}\right]} + \sqrt{TK\mathbb{E}\left[\sum_{\ell}\sum_{i\in S_\ell}\frac{\log^2 TK}{n_i^2(\ell)}\right]}\right),
\end{aligned} \tag{20}$$

where the last inequality follows Cauchy-Schwarz inequality. Since $v_i\leq 1$ for all i, we have,

$$\sum_{\ell}\sum_{i\in S_\ell}\frac{v_i\log TK}{n_i(\ell)} \leq \sum_{i=1}^{N}\sum_{n_i(\ell)=1}^{T_i}\frac{\log TK}{n_i(\ell)} \leq N\log TK\cdot\log T,$$

and

$$\sum_{\ell}\sum_{i\in S_\ell}\frac{\log^2 TK}{n_i^2(\ell)} = \sum_{i=1}^{N}\sum_{n_i(\ell)=1}^{T_i}\frac{\log^2 TK}{n_i(\ell)} \leq 4N\log^2 TK,$$

Therefore by substituting preceeding two inequalities in (20), we obtain that

$$\mathsf{Reg}_1(T,\mathbf{v}) \leq C\sqrt{\frac{NT}{K}},$$

for some constant C. The result follows from this inequality and (16).

5. Empirical study In this section, we test the various design components of our Thompson Sampling approach through numerical simulations. The aim is to isolate and understand the effect of individual features of our algorithm like Beta posteriors vs. Gaussian approximation, independent sampling vs. correlated sampling, and single sample vs. multiple samples, on the practical performance.

We simulate an instance of the MNL-Bandit problem with $N = 1000$, $K = 10$ and $T = 2\times 10^5$, when the MNL parameters $\{v_i\}_{i=1,\ldots,N}$ are generated randomly from $\mathsf{Unif}[0,1]$. And, we compute the average regret based on 50 independent simulations over the randomly generated instance. In Figure 1, we report performance of following different variants of TS:

$i)$ Algorithm 1: Thompson Sampling with independent Beta priors, as described in Algorithm 1.

$ii)$ $\mathsf{TS}_{\mathsf{IID\ Gauss}}$: Algorithm 1 with Gaussian posterior approximation and independent sampling. More specifically, for each epoch ℓ and for each item i, we sample a Gaussian random variable independently with the mean and variance equal to the mean and variance of the Beta prior in Algorithm 1 (see Lemma 6).

$iii)$ $\mathsf{TS}_{\mathsf{Gauss\ Corr}}$: Algorithm 1 with Gaussian posterior approximation and correlated sampling. In particular, for every epoch ℓ, we sample a standard normal random variable. Then for each item i, we obtain a corresponding sample by multiplying and adding the preceding sample with the standard deviation and mean of the Beta prior in Algorithm 1 (see Step (a) in Algorithm 2). We use the values $\alpha=\beta=1$ for this variant of Thompson Sampling.

$iv)$ Algorithm 2: Algorithm 1 with Gaussian posterior approximation with correlated sampling and boosting by using multiple (K) samples. This is essentially the version with all the features of Algorithm 2. We use the values $\alpha=\beta=1$ for this variant of Thompson Sampling.

For comparison, we also present the performance of UCB approach presented in [2]. The performance of all the variants of TS is observed to be better than the UCB approach in our experiments, which is consistent with the other empirical evidence in the literature.

Figure 1 shows the performance of the TS variants. Among the TS variants, the performance of Algorithm 1, i.e., Thompson Sampling with independent Beta priors is similar to $\mathsf{TS}_{\mathsf{IID\ Gauss}}$, the version with independent Gaussian (approximate) posteriors; indicating that the effect of posterior approximation is minor. The performance of $\mathsf{TS}_{\mathsf{Gauss\ Corr}}$, where we generate correlated samples from the Gaussian distributions, is significantly better than the other variants of the algorithm. This is consistent with our remark earlier that to adapt the Thompson sampling approach of the classical MAB problem to our setting, ideally, we would like to maintain a joint prior over the parameters $\{v_i\}_{i=1,\ldots,N}$ and update it to a joint posterior using the Bandit feedback. However, since this can be quite challenging, and intractable in general, we use independent priors over the parameters. The superior performance of $\mathsf{TS}_{\mathsf{Gauss\ Corr}}$ demonstrates the potential benefits of considering a joint (correlated) prior/posterior in settings with a combinatorial structure. Finally, we observe that the performance of Algorithm 2, where an additional "variance boosting" is provided through K independent samples, is worse than $\mathsf{TS}_{\mathsf{Gauss\ Corr}}$. Note that while "variance boosting" facilitates theoretical analysis, it also results in a longer exploration period explaining the observed degradation of performance in comparison to the TS variant without "variance boosting." However, Algorithm 2 performance significantly better than the independent Beta posterior version Algorithm 1. Therefore, significant improvements in performance due to the correlated sampling feature of Algorithm 2 compensate for the slight deterioration caused by boosting.

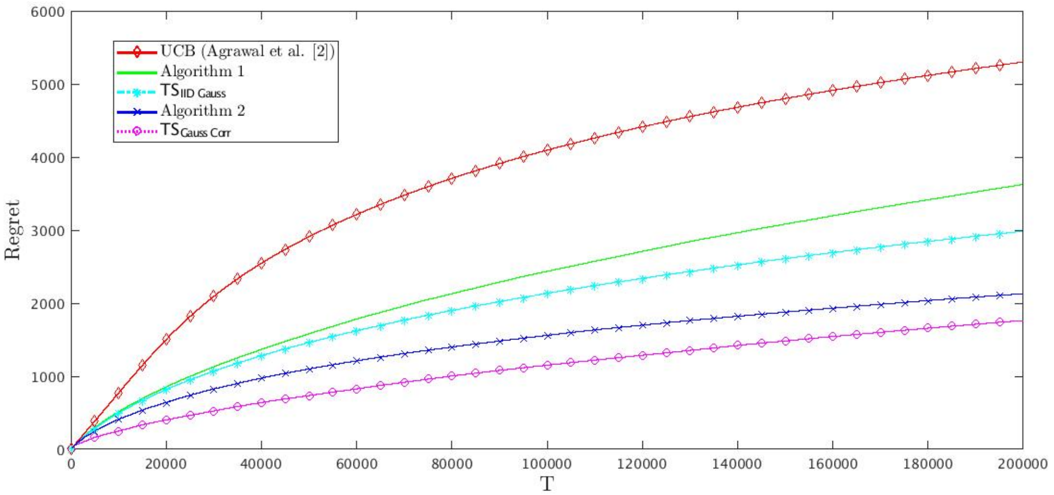

FIGURE 1. Regret growth with T for various heuristics on a randomly generated MNL-Bandit instance with $N = 1000, K = 10$.

6. Conclusion. In this paper, we consider a combinatorial variant of the traditional multi-armed Bandit problem, MNL-Bandit, and present a TS-based policy for this problem. Focusing on designing a computationally efficient algorithm that facilitates theoretical analysis, we highlight several challenges involved in adaptive TS-based approaches for the MNL-Bandit problem and discuss algorithm design choices to address them. To the best of our knowledge, the idea of correlated sampling for combinatorial arms is novel and potentially useful for further combinatorial dynamic learning.

Appendix A: Unbiased Estimate $\tilde{v}_{i,\ell}$ and Conjugate priors Lemma 1 establishes that the estimate obtained from epoch based offerings, $\tilde{v}_{i,\ell}$ in Algorithm 1 is an unbiased estimate and is distributed geometrically with probability of success $\frac{1}{v_i+1}$. This result is adapted from Agrawal et al. [2] and we provide the proof for completeness.

Proof of Lemma 1: We prove the result by computing the moment generating function, from which we can establish that $\tilde{v}_{i,\ell}$ is a geometric random variable with parameter $\frac{1}{1+v_i}$. More specifically, we show that the moment generating function of estimate conditioned on S_ℓ, $\hat{v}_i$, is given by,

$$\mathbb{E}\left(e^{\theta\tilde{v}_{i,\ell}}\middle|S_\ell\right) = \frac{1}{1-v_i(e^\theta-1)}, \text{ for all } \theta \le \log\frac{1+v_i}{v_i}, \text{ for all } i = 1,\cdots,N,$$

thereby also establishing that $\tilde{v}_{i,\ell}$ are unbiased estimators of v_i.

We focus on proving the above result. From (1), we have that the probability of the no purchase event when assortment S_ℓ is offered is given by

$$p_0(S_\ell) = \frac{1}{1+\sum_{j\in S_\ell} v_j}.$$

Let n_ℓ be the total number of offerings in epoch ℓ before a no purchased occurred, i.e., $n_\ell = |\mathcal{E}_\ell| - 1$. Therefore, n_ℓ is a geometric random variable with probability of success $p_0(S_\ell)$. And, given any fixed value of n_ℓ, $\tilde{v}_{i,\ell}$ is a binomial random variable with n_ℓ trials and probability of success given by

$$q_i(S_\ell) = \frac{v_i}{\sum_{j\in S_\ell} v_j}.$$

In the calculations below, for brevity we use p_0 and q_i respectively to denote $p_0(S_\ell)$ and $q_i(S_\ell)$. Hence, we have

$$\mathbb{E}\left(e^{\theta\tilde{v}_{i,\ell}}\right) = E_{n_\ell}\left\{\mathbb{E}\left(e^{\theta\tilde{v}_{i,\ell}}\,\middle|\,n_\ell\right)\right\}.$$

Since the moment generating function for a binomial random variable with parameters n, p is $(pe^\theta + 1 - p)^n$, we have

$$\mathbb{E}\left(e^{\theta \tilde{v}_{i,\ell}} \,\middle|\, n_\ell\right) = E_{n_\ell}\left\{\left(q_i e^\theta + 1 - q_i\right)^{n_\ell}\right\}.$$

For any α, such that, $\alpha(1-p) < 1$ n is a geometric random variable with parameter p, we have

$$\mathbb{E}(\alpha^n) = \frac{p}{1 - \alpha(1-p)}.$$

Note that for all $\theta < \log \frac{1+v_i}{v_i}$, we have $\left(q_i e^\theta + (1-q_i)\right)(1-p_0) = (1-p_0) + p_0 v_i(e^\theta - 1) < 1$. Therefore, we have

$$\mathbb{E}\left(e^{\theta \tilde{v}_{i,\ell}}\right) = \frac{1}{1 - v_i(e^\theta - 1)} \text{ for all } \theta < \log \frac{1+v_i}{v_i}.$$

This concludes the proof. □

Building on this result. We will prove Lemma 2 that helped construct Algorithm 1. Recall in Lemma 2, we show that the distribution of $\tilde{v}_{i,\ell}$ has a conjugate prior.

Proof of Lemma 2: The proof of the lemma follows from the following result on the probability density function of the random variable $X_{\alpha,\beta}$. Specifically, we have for any $x > 0$

$$f_{\alpha,\beta}(x) = \frac{1}{B(\alpha,\beta)}\left(\frac{1}{1+x}\right)^{\alpha+1}\left(\frac{x}{x+1}\right)^{\beta-1}, \tag{21}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a)$ is the gamma function. Since we assume that the parameter v_i's prior distribution is same as that of $X_{\alpha,\beta}$, we have from (21) and Lemma 1,

$$\mathbb{P}\left(v_i \middle| \tilde{v}_{i,\ell} = m\right) \propto \left(\frac{1}{1+v_i}\right)^{\alpha+2}\left(\frac{v_i}{v_i+1}\right)^{\beta+m-1}.$$

□

Given the pdf of the posterior in (21), it is possible to compute the mean and variance of the posterior distribution. We show that they have simple closed form expressions. Now we will prove Lemma 3, which provides the moments of the aforementioned posterior distribution.

Proof of Lemma 3 We prove the result by relating the mean of the posterior to the mean of the Beta distribution. Let $\hat{X} = \frac{1}{X} - 1$. From (21), we have

$$\mathbb{E}(\hat{X}) = \frac{1}{B(\alpha,\beta)}\int_0^\infty x\left(\frac{1}{1+x}\right)^{\alpha+1}\left(\frac{x}{x+1}\right)^{\beta-1} dx,$$

Substituting $y = \frac{1}{1+x}$, we have

$$\mathbb{E}(\hat{X}) = \frac{1}{B(\alpha,\beta)}\int_0^1 y^{\alpha-2}(1-y)^\beta dx = \frac{B(\alpha-1,\beta+1)}{B(\alpha,\beta)} = \frac{\beta}{\alpha-1}.$$

Similarly, we can derive the expression for the $\mathsf{Var}(\hat{X})$. This concludes the proof. □

A.1. Some concentration bounds In this section, we prove bounds on how fast our estimate $\hat{v}_i$ converges to the true mean The concentration bounds we prove in the section are similar to Chernoff bounds, but for the fact that $n_i(\ell)$ is a random variable. Hence, we use a self-normalized martingale technique to derive concentration bounds. We will then utilize the large deviation properties of Gaussian distribution to show that the posterior distributions concentrate around their means.

Lemma 9 *Let δ_i, $i = 1, \cdots, N$ be arbitrary random variables. If $v_i \leq 1$, for all $i = 1, \cdots, N$, then we have, for all $i = 1, \cdots, N$,*

1.

$$\mathbb{P}\left(\hat{v}_i(\ell) > (1+\delta_i)v_i\right) \leq \left(\mathbb{E}\left[\exp\left(-\frac{v_i\delta_i^2 n_i(\ell)}{2(1+\delta_i)(1+v_i)^2}\right)\right]\right)^{\frac{1}{2}},$$

and

2.

$$\mathbb{P}\left(\hat{v}_i(\ell) < (1-\delta_i)v_i\right) \leq \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)^2}\left(3 - \frac{2\delta_i v_i}{1+v_i}\right)\right)\right].$$

Proof. Fix $i \in \{1, ldots, n\}$. We have

$$\hat{v}_i(\ell) = \frac{1}{n_i(\ell)}\sum_{\tau=1}^{\ell} \tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau).$$

Therefore, bounding $\mathbb{P}\left(\hat{v}_i(\ell) > (1+\delta_i)v_i\right)$ and $\mathbb{P}\left(\hat{v}_i(\ell) < (1-\delta_i)v_i\right)$ is equivalent to bounding $\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) > (1+\delta)v_i n_i(\ell)\right)$ and $\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) < (1-\delta)v_i n_i(\ell)\right)$. We will bound the first term and then follow a similar approach for bounding the second term to complete the proof.

Bounding $\mathbb{P}\left(\hat{v}_i(\ell) > (1+\delta_i)v_i\right)$: From Markov Inequality, we have for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) > (1+\delta_i)v_i n_i(\ell)\right) &= \mathbb{P}\left\{\exp\left(\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau)\right) > \exp\left(\lambda(1+\delta_i)v_i n_i(\ell)\right)\right\}, \\
&= \mathbb{P}\left\{\exp\left(\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) - \lambda(1+\delta_i)v_i n_i(\ell)\right) > 1\right\}, \qquad (22) \\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) - \lambda(1+\delta_i)v_i n_i(\ell)\right)\right].
\end{aligned}$$

For notational brevity, denote by $f(\lambda, v_i)$ the function,

$$f(\lambda, v_i) = -\frac{\log\left(1 - v_i(e^{2\lambda} - 1)\right)}{2}.$$

We have,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau) - \lambda(1+\delta_i)v_i n_i(\ell)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(\lambda\tilde{v}_{i,\tau} - f(\lambda, v_i))\cdot\mathbb{1}(i \in S_\tau)\right)\cdot\exp\Big(-\lambda(1+\delta_i)v_i(1 - f(\lambda, v_i))n_i(\ell)\Big)\right], \qquad (23) \\
&\leq \left(\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(2\lambda\tilde{v}_{i,\tau} - 2f(\lambda, v_i))\cdot\mathbb{1}(i \in S_\tau)\right)\right]\cdot\mathbb{E}\Big[\exp\Big(-2\lambda(1+\delta_i)v_i(1 - f(\lambda, v_i))n_i(\ell)\Big)\Big]\right)^{\frac{1}{2}},
\end{aligned}$$

where the above follows from Cauchy-Schwartz inequality. Let $\mathcal{F}_\tau$ be the filtration corresponding to the history until epoch τ. Note that for any τ, $\mathbb{1}(i \in S_\tau)$ is measurable on $\mathcal{F}_\tau$ and $\{\tilde{v}_{i,\tau}|\mathcal{F}_\tau\}$ is a geometric random variable. From the proof of Lemma 1, for all $\tau \geq 1$ and for any $0 < \lambda < \frac{1}{2}\log\frac{1+v_i}{v_i}$, we have,

$$\mathbb{E}\left(e^{2\lambda\tilde{v}_{i,\tau}\mathbb{1}(i \in S_\tau)}\middle|\mathcal{F}_\tau\right) = \left(\frac{1}{1 - v_i(e^{2\lambda} - 1)}\right)^{\mathbb{1}(i \in S_\tau)}.$$

Therefore, it follows that

$$\mathbb{E}\left(e^{(2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\cdot\mathbb{1}(i\in S_\tau)}\middle|\mathcal{F}_\tau\right)\leq 1, \tag{24}$$

and

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left\{\exp\left((2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)|\mathcal{F}_\ell\right\}\right]\\ &= \mathbb{E}\left[\prod_{\tau=1}^{\ell-1}\exp\left((2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\mathbb{E}\left(e^{(2\lambda\tilde{v}_{i,\ell}-2f(\lambda,v_i))\mathbb{1}(i\in S_\ell)}\middle|\mathcal{F}_\ell\right)\right]\\ &\leq \mathbb{E}\left[\prod_{\tau=1}^{\ell-1}\exp\left((2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right],\end{aligned}$$

where the inequality follows from (24). Similarly by conditioning with $\mathcal{F}_{\ell-1},\cdots,\mathcal{F}_1$, we obtain,

$$\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right]\leq 1.$$

From (22) and (23), we have

$$\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)>(1+\delta_i)v_in_i(\ell)\right)\leq\left(\mathbb{E}\left[\exp\Big(-2\lambda(1+\delta_i)v_i(1-f(\lambda,v_i))n_i(\ell)\Big)\right]\right)^{\frac{1}{2}}.$$

Therefore, we have

$$\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)>(1+\delta_i)v_in_i(\ell)\right)\leq\left(\mathbb{E}\left[\min_{\lambda\in\Omega}\exp\Big(-2\lambda(1+\delta_i)v_i(1-f(\lambda,v_i))n_i(\ell)\Big)\right]\right)^{\frac{1}{2}}, \tag{25}$$

where $\Omega=\{\lambda\mid 0<\lambda<\frac{1}{2}\log\frac{1+v_i}{v_i}\}$ is the range of λ for which the moment generating function in (24) is well definred. Taking logarithm of the objective in (25), we have,

$$\operatorname*{argmin}_{\lambda\in\Omega}\left\{e^{-2\lambda(1+\delta_i)v_i(1-f(\lambda,v_i))\cdot n_i(\ell)}=\operatorname*{argmin}_{\lambda\in\Omega}-2(1+\delta_i)\lambda n_i(\ell)v_i-n_i(\ell)\log\left(1-v_i(e^{2\lambda}-1)\right)\right\}. \tag{26}$$

Noting that the right hand side in the above equation is a convex function in λ, we obtain the optimal λ by solving for the first order conditions. Specifically, at optimal t, we have

$$e^{2\lambda}=\frac{(1+\delta_i)(1+v_i)}{1+v_i(1+\delta_i)}.$$

Substituting the above expression in (25), we obtain the following bound.

$$\mathbb{P}\left(\hat{v}_i(\ell)>(1+\delta_i)v_i\right)\leq\mathbb{E}^{\frac{1}{2}}\left[\left(1-\frac{\delta_i}{(1+\delta_i)(1+v_i)}\right)^{n_i(\ell)v_i(1+\delta_i)}\left(1+\frac{\delta_iv_i}{1+v_i}\right)^{n_i(\ell)}\right]. \tag{27}$$

For notational brevity, we will use n to denote the random variable $n_i(\ell)$ and focus on bounding the right hand term in the above equation. From Taylor series of $\log(1-x)$, we have that

$$nv_i(1+\delta_i)\log\left(1-\frac{\delta_i}{(1+\delta_i)(1+v_i)}\right)\leq-\frac{n\delta_iv_i}{1+v_i}-\frac{n\delta_i^2v_i}{2(1+\delta_i)(1+v_i)^2},$$

and similarly for $\log(1+x)$, we have

$$n\log\left(1+\frac{\delta_i v_i}{1+v_i}\right) \le \frac{n\delta_i v_i}{(1+v_i)}.$$

Note that if $\delta_i > 1$, we can use the fact that $\log(1+\delta_i x) \le \delta_i \log(1+x)$ to arrive at the preceding result. Substituting the preceding two equations in (27), we have

$$\mathbb{P}\left(\hat{v}_i(\ell) > (1+\delta_i)v_i\right) \le \mathbb{E}^{\frac{1}{2}}\left[\exp\left(-\frac{n\delta_i^2 v_i}{2(1+\delta_i)(1+v_i)^2}\right)\right]. \tag{28}$$

Bounding $\mathbb{P}\left(\hat{v}_i(\ell) < (1-\delta_i)v_i\right)$: Now to bound the other one sided inequality, we use the fact that for any $\lambda > 0$,

$$\mathbb{E}\left(e^{-\lambda\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)}\middle|\mathcal{F}_\tau\right) = \left(\frac{1}{1-v_i(e^{-\lambda}-1)}\right)^{\mathbb{1}(i\in S_\tau)}.$$

and follow a similar approach. More specifically, from Markov inequality, for any $\lambda > 0$ and $0 < \delta_i < 1$, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau) < (1-\delta_i)v_i n_i(\ell)\right) &= \mathbb{P}\left\{\exp\left(-\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)\right) > \exp\left(-\lambda(1-\delta_i)v_i n_i(\ell)\right)\right\},\\ &= \mathbb{P}\left\{\exp\left(-\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau) + \lambda(1-\delta_i)v_i n_i(\ell)\right) > 1\right\}, \quad (29)\\ &\le \mathbb{E}\left[\exp\left(-\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau) + \lambda(1-\delta_i)v_i n_i(\ell)\right)\right].\end{aligned}$$

For notational brevity, denote $f(\lambda, v_i)$ by the function,

$$f(\lambda, v_i) = -\frac{\log\left(1-v_i(e^{-2\lambda}-1)\right)}{2}.$$

We have,

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\lambda\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau) + \lambda(1-\delta_i)v_i n_i(\ell)\right)\right]\\ &= \mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(-\lambda\tilde{v}_{i,\tau} - f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\exp\Big(\lambda(1-\delta_i)v_i(1+f(\lambda,v_i))n_i(\ell)\Big)\right], \quad (30)\\ &\le \left(\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(-2\lambda\tilde{v}_{i,\tau} - 2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right]\mathbb{E}\Big[\exp\Big(2\lambda(1-\delta_i)v_i(1+f(\lambda,v_i))n_i(\ell)\Big)\Big]\right)^{\frac{1}{2}},\end{aligned}$$

where the above inequality follows from Cauchy-Schwartz inequality. Let $\mathcal{F}_\tau$ be the filtration corresponding to the history until epoch τ. Note that for any τ, $\mathbb{1}(i\in S_\tau)$ conditioned on F_τ is a constant and $\{\tilde{v}_{i,\tau}|\mathcal{F}_\tau\}$ is a geometric random variable. Therefore, for all $\tau \ge 1$ and for any $\lambda > 0$, we have,

$$\mathbb{E}\left(e^{-2\lambda\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)}\middle|\mathcal{F}_\tau\right) = \left(\frac{1}{1-v_i(e^{-2\lambda}-1)}\right)^{\mathbb{1}(i\in S_\tau)}.$$

Therefore, it follows that

$$\mathbb{E}\left(e^{(-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)}\middle|\mathcal{F}_\tau\right) \le 1, \tag{31}$$

and

$$\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right]=\mathbb{E}\left[\mathbb{E}\left\{\exp\left(\sum_{\tau=1}^{\ell}(-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\middle|\mathcal{F}_\ell\right\}\right],$$
$$=\mathbb{E}\left[\prod_{\tau=1}^{\ell-1}\exp\left((-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\mathbb{E}\left(e^{(-2\lambda\tilde{v}_{i,\ell}-2f(\lambda,v_i))\mathbb{1}(i\in S_\ell)}\middle|\mathcal{F}_\ell\right)\right],$$
$$=\mathbb{E}\left[\prod_{\tau=1}^{\ell-1}\exp\left((-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right],$$

where the inequality follows from (31). Similarly by conditioning with $\mathcal{F}_{\ell-1},\cdots,\mathcal{F}_1$, we obtain,

$$\mathbb{E}\left[\exp\left(\sum_{\tau=1}^{\ell}(-2\lambda\tilde{v}_{i,\tau}-2f(\lambda,v_i))\mathbb{1}(i\in S_\tau)\right)\right]\leq 1.$$

From (29) and (30), we have

$$\mathbb{P}\left(\sum_{\tau=1}^{\ell}\tilde{v}_{i,\tau}\mathbb{1}(i\in S_\tau)<(1-\delta_i)v_i n_i(\ell)\right)\leq\left(\mathbb{E}\left[\exp\left(2\lambda(1-\delta_i)v_i(1+f(\lambda,v_i))n_i(\ell)\right)\right]\right)^{\frac{1}{2}}.$$

Therefore, we have

$$\mathbb{P}\left(\hat{v}_i(\ell)<(1-\delta_i)v_i\right)\leq\left(\mathbb{E}\left[\min_{\lambda>0}\exp\left(2\lambda(1-\delta_i)v_i(1+f(\lambda,v_i))n_i(\ell)\right)\right]\right)^{\frac{1}{2}}.$$

Following similar approach as in optimizing the previous bound (see (25)) to establish the following result. For notational brevity, we will use n to denote the random variable $n_i(\ell)$.

$$\mathbb{P}\left(\hat{v}_i(\ell)<(1-\delta_i)v_i\right)\leq\mathbb{E}^{\frac{1}{2}}\left[\left(1+\frac{\delta_i}{(1-\delta_i)(1+v_i)}\right)^{nv_i(1-\delta_i)}\left(1-\frac{\delta_i v_i}{1+v_i}\right)^{n}\right].$$

Now we will use Taylor series for $\log(1+x)$ and $\log(1-x)$ in a similar manner as described for the other bound to obtain the required result. In particular, since $1-\delta_i\leq 1$, we have for any $x>0$ it follows that $(1+\frac{x}{1-\delta_i})^{(1-\delta_i)}\leq(1+x)$. Therefore, we have

$$\mathbb{P}\left(\hat{v}_i(\ell)<(1-\delta_i)v_i\right)\leq\left(\mathbb{E}\left[\left(1+\frac{\delta_i}{1+v_i}\right)^{nv_i}\left(1-\frac{\delta_i v_i}{1+v_i}\right)^{n}\right]\right)^{\frac{1}{2}}. \tag{32}$$

Note that since $\tilde{v}_{i,\tau}\geq 0$ for all i,τ, we have a zero probability event if $\delta_i>1$. Therefore, without loss of generality, we assume $\delta_i<1$ and from Taylor series for $\log(1-x)$, we have

$$n\log\left(1-\frac{\delta_i v_i}{1+v_i}\right)\leq-\frac{n\delta_i v_i}{1+v_i},$$

and from Taylor series for $\log(1+x)$, we have

$$n\log\left(1+\frac{\delta_i v_i}{1+v_i}\right)\leq\frac{n\delta_i}{(1+v_i)}-\frac{n\delta_i^2 v_i}{6(1+v_i)^2}\left(3-\frac{2\delta_i v_i}{1+v_i}\right).$$

Therefore, substituting the preceding equations in (32), we have,

$$\mathbb{P}\left(\hat{v}_i<(1-\delta_i)v_i\right)\leq\exp\left(-\frac{n\delta_i^2 v_i}{6(1+v_i)^2}\left(3-\frac{2\delta_i\mu}{1+v_i}\right)\right). \tag{33}$$

The result follows from (28) and (33). □

Proof of Lemma 6. Let $\delta_i = \sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}}$. We analyze the cases $\delta_i \leq \frac{1}{2}$ and $\delta_i \geq \frac{1}{2}$ separately.

Case 1: $\delta_i \leq \frac{1}{2}$: For any $v_i \leq 1$ and $\delta_i \leq 1/2$, we have,

$$\frac{v_i\delta_i^2 n_i(\ell)}{2(1+\delta_i)(1+v_i)^2} \geq \frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)} \geq m\log(\rho+1),$$

and

$$\frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)^2}\left(3-\frac{2\delta_i v_i}{1+v_i}\right) \geq \frac{v_i\delta_i^2 n_i(\ell)}{6(1+v_i)} \geq m\log(\rho+1).$$

Therefore, substituting $\delta_i = \sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}}$ in Lemma 9, we have,

$$\begin{aligned}\mathcal{P}\left(2\hat{v}_i(\ell) \geq v_i\right) &\geq 1-\frac{1}{\rho^m},\\ \mathbb{P}\left\{|\hat{v}_i(\ell)-v_i| < \sqrt{\frac{4v_i(v_i+2)m\log(\rho+1)}{n_i(\ell)}}\right\} &\geq 1-\frac{2}{\rho^m}.\end{aligned} \tag{34}$$

From the above three results, we have,

$$\mathbb{P}\left\{|\hat{v}_i(\ell)-v_i| < \sqrt{\frac{16\hat{v}_i(\ell)\left(\hat{v}_i(\ell)+1\right)\log(\rho+1)}{n_i(\ell)}}\right\} \geq \mathbb{P}\left\{|\hat{v}_i(\ell)-v_i| < \sqrt{\frac{4v_i(v_i+2)\log(\rho+1)}{n_i(\ell)}}\right\} \geq 1-\frac{3}{\rho^m}. \tag{35}$$

By assumption, $v_i \leq 1$. Therefore, we have $v_i(v_i+2) \leq 3v_i$ and,

$$\mathbb{P}\left\{|\hat{v}_i(\ell)-v_i| < \sqrt{\frac{12v_i\log(\rho+1)}{n_i(\ell)}}\right\} \geq 1-\frac{3}{\rho^m}.$$

Case 2: $\delta_i > \frac{1}{2}$: Now consider the scenario, when $\sqrt{\frac{4(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)}} > \frac{1}{2}$. Then, we have,

$$\bar{\delta}_i := \frac{8(v_i+2)m\log(\rho+1)}{v_i n_i(\ell)} \geq \frac{1}{2},$$

which implies for any $v_i \leq 1$,

$$\begin{aligned}\frac{nv_i\bar{\delta}_i^2}{2(1+\bar{\delta}_i)(1+v_i)^2} &\geq \frac{nv_i\bar{\delta}_i}{12(1+v_i)},\\ \frac{n\bar{\delta}_i^2 v_i}{6(1+v_i)^2}\left(3-\frac{2\bar{\delta}_i v_i}{1+v_i}\right) &\geq \frac{nv_i\bar{\delta}_i}{12(1+v_i)}.\end{aligned}$$

Therefore, substituting the value of $\bar{\delta}_i$ in Lemma 9, we have

$$\mathbb{P}\left\{|\hat{v}_i(\ell)-v_i| > \frac{24m\log(\rho+1)}{n}\right\} \leq \frac{2}{\rho^m}.$$

This completes the proof. □

Proof of Lemma 5: Note that we have $\mu_i(\ell) = \hat{v}_i(\ell) + \hat{\sigma}_i(\ell)\cdot \max_{j=1,\cdots,K}\{\theta^{(j)}(\ell)\}$. Therefore, from union bound, we have,

$$\begin{aligned}\mathbb{P}\left\{|\mu_i(\ell)-\hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\;\middle|\hat{v}_i(\ell)\right\} &= \mathbb{P}\left(\bigcup_{j=1}^{K}\left\{\theta^j(\ell) > 4\sqrt{\log rK}\right\}\right)\\ &\leq \sum_{j=1}^{K}\mathbb{P}\left(\theta^j(\ell) > 4\sqrt{\log rK}\right)\end{aligned}$$

The result follows from the above inequality and the following anti-concentration bound for the normal random variable $\theta^{(j)}(\ell)$ (see formula 7.1.13 in Abramowitz and Stegun [1]).

$$\frac{1}{4\sqrt{\pi}} \cdot e^{-7z^2/2} < \mathbb{P}\left(|\theta^{(j)}(\ell)| > z\right) \leq \frac{1}{2} e^{-z^2/2}.$$

□

Corollary A.1 *For any item i and any epoch ℓ, we have*

$$\mathbb{E}\left(|\mu_i(\ell) - \hat{v}_i(\ell)|\right) \leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK}.$$

Proof. In Lemma 5, we show that for any $r > 0$ and $i = 1, \cdots, N$, we have,

$$\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right) \leq \frac{1}{r^4 K^3},$$

where $\hat{\sigma}_i(\ell) = \sqrt{\frac{50\hat{v}_i(\hat{v}_i+1)}{n_i}} + \frac{75\sqrt{\log TK}}{n_i}$. Since $S_\ell \subset \{1, \cdots, N\}$, we have for any $i \in S_\ell$ and $r > 0$, we have

$$\begin{aligned}\mathbb{P}\left(|\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK} \text{ for any } i \in S_\ell\right) &\leq \mathbb{P}\left(\bigcup_{i=1}^{N} |\mu_i(\ell) - \hat{v}_i(\ell)| > 4\hat{\sigma}_i(\ell)\sqrt{\log rK}\right), \\ &\leq \frac{N}{r^4 K^3}.\end{aligned} \tag{36}$$

Since $|\mu_i(\ell) - \hat{v}_i(\ell)|$ is a non-negative random variable, we have

$$\begin{aligned}\mathbb{E}(|\mu_i(\ell) - \hat{v}_i(\ell)|) &= \int_0^\infty \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx, \\ &= \int_0^{4\hat{\sigma}_i(\ell)\sqrt{\log TK}} \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx + \int_{4\hat{\sigma}_i(\ell)\sqrt{\log TK}}^\infty \mathbb{P}\left\{|\mu_i(\ell) - \hat{v}_i(\ell)| \geq x\right\} dx, \\ &\leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK} + \sum_{r=T}^{\infty} \int_{4\hat{\sigma}_i(\ell)\sqrt{\log rK}}^{4\hat{\sigma}_i(\ell)\sqrt{\log (r+1)K}} \mathbb{P}\left\{Y \geq x\right\} dx, \\ &\overset{(a)}{\leq} 4\hat{\sigma}_i(\ell)\sqrt{\log TK} + \sum_{r=T}^{\infty} \frac{N\sqrt{\log (rK+1)} - N\sqrt{\log rK}}{r^4 K^3}, \\ &\leq 4\hat{\sigma}_i(\ell)\sqrt{\log TK} \;\text{ for any } T \geq N,\end{aligned} \tag{37}$$

where the inequality (a) follows from (36).

Acknowledgments. Shipra Agrawal is supported in part by NSF Grant CMMI-1846792 (CAREER) . V. Goyal is supported in part by NSF Grants CMMI-1351838 (CAREER) and CMMI-1636046. A. Zeevi is supported in part by NSF Grants NetSE-0964170 and BSF-2010466.

References

[1] M. Abramowitz and I. A. Stegun. 1964. Handbook of mathematical functions: with formulas, graphs, and mathematical tables. (1964).

[2] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. 2016. A Near-Optimal Exploration-Exploitation Approach for Assortment Selection. *Proceedings of the 2016 ACM Conference on Economics and Computation (EC)* (2016), 599–600.

[3] S. Agrawal and N. Goyal. 2013a. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. (31). 99–107.

[4] S. Agrawal and N. Goyal. 2013b. Thompson Sampling for Contextual Bandits with Linear Payoffs.. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, Vol. (28). 127–135.

[5] P. Auer. 2003. Using Confidence Bounds for Exploitation-exploration Trade-offs. *Journal of Machine Learning Research* (3) (2003), 397–422.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47 (2002), 235–256.

[7] V. Avadhanula, J. Bhandari, V. Goyal, and A. Zeevi. 2016. On the tightness of an LP relaxation for rational optimization and its applications. *Operations Research Letters* 44, (5) (2016), 612–617.

[8] M. Ben-Akiva and S. Lerman. 1985. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press.

[9] X. Chen, A. Krishnamurthy, and Y. Wang. 2019. Robust Dynamic Assortment Optimization in the Presence of Outlier Customers. *arXiv preprint arXiv:1910.04183* (2019).

[10] X. Chen, Y. Wang, and Y. Zhou. 2018a. Dynamic assortment optimization with changing contextual information. *arXiv preprint arXiv:1810.13069* (2018).

[11] X. Chen, Y. Wang, and Y. Zhou. 2018b. Dynamic assortment selection under the nested logit models. *arXiv preprint arXiv:1806.10410* (2018).

[12] W. Cheung and D. Simchi-Levi. 2017. Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658* (2017).

[13] J. Davis, G. Gallego, and H. Topaloglu. 2013. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Technical Report* (2013).

[14] Y. Feng, R. Caldentey, and Christopher T. R. 2018. Learning Customer Preferences from Personalized Assortments. *Available at SSRN 3215614* (2018).

[15] A. Gopalan, S. Mannor, and Y. Mansour. 2014. Thompson Sampling for Complex Online Problems.. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, Vol. (32). 100–108.

[16] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. 2010. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th international conference on machine learning (ICML)*. 13–20.

[17] R.D. Luce. 1959. *Individual choice behavior: A theoretical analysis*. Wiley.

[18] B. C. May, N. Korda, A. Lee, and D. S. Leslie. 2012. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* (13) (2012), 2069–2106.

[19] D. McFadden. 1978. *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.

[20] S. Miao and X. Chao. 2018. Dynamic Joint Assortment and Pricing Optimization With Demand Learning. *Available at SSRN 3173267* (2018).

[21] S. Miao and X. Chao. 2019. Fast Algorithms for Online Personalized Assortment Optimization in a Big Data Regime. *Available at SSRN 3432574* (2019).

[22] M. Oh and G. Iyengar. 2019a. Multinomial Logit Contextual Bandits. (2019).

[23] M. Oh and G. Iyengar. 2019b. Thompson Sampling for Multinomial Logit Contextual Bandits. In *Advances in Neural Information Processing Systems*. 3145–3155.

[24] C. Oliver and L. Li. 2011. An Empirical Evaluation of Thompson Sampling. *In Advances in Neural Information Processing Systems (NIPS)* 24 (2011), 2249?2257.

[25] R. L. Plackett. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* (1975).

[26] P. Rusmevichientong, Z. M. Shen, and D.B. Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research* 58, (6) (2010), 1666–1680.

[27] P. Rusmevichientong and J.N. Tsitsiklis. 2010. Linearly Parameterized Bandits. *Mathematics of Operations of Research* 35(2) (2010), 395–411.

[28] D. Russo and B. Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39, 4 (2014), 1221–1243.

[29] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, and others. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.

[30] A. Saha and A. Gopalan. 2019. Regret Minimisation in Multinomial Logit Bandits. *arXiv preprint arXiv:1903.00543* (2019).

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

[32] W.R. Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[33] K. Train. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.

[34] Y. Wang, X. Chen, and Y. Zhou. 2018. Near-optimal policies for dynamic multinomial logit assortment selection models. In *Advances in Neural Information Processing Systems*. 3101–3110.