Sequential Learning of Product Recommendations with Customer Disengagement

Hamsa Bastani
Wharton School, Operations Information and Decisions, hamsab@wharton.upenn.edu

Pavithra Harsha
IBM Thomas J. Watson Research, pharsha@us.ibm.com

Georgia Perakis
MIT Sloan School of Management, Operations Management, georgiap@mit.edu

Divya Singhvi
MIT Operations Research Center, dsinghvi@mit.edu

We consider the problem of sequential product recommendation when customer preferences are unknown. First, we present empirical evidence of customer disengagement using a sequence of ad campaigns from a major airline carrier. In particular, customers decide to stay on the platform based on the relevance of recommendations. We then formulate this problem as a linear bandit, with the notable difference that the customer's horizon length is a function of past recommendations. We prove that any algorithm in this setting achieves linear regret. Thus, no algorithm can keep all customers engaged; however, we can hope to keep a subset of customers engaged. Unfortunately, we find that classical bandit learning as well as greedy algorithms provably over-explore, thereby incurring linear regret for every customer. We propose modifying bandit learning strategies by constraining the action space upfront using an integer program. We prove that this simple modification allows our algorithm to achieve sublinear regret for a significant fraction of customers. Furthermore, numerical experiments on real movie recommendations data demonstrate that our algorithm can improve customer engagement with the platform by up to 80%.

Key words: bandits, online learning, recommendation systems, disengagement, cold start
History: This paper is under preparation.

1. Introduction

Personalized customer recommendations are a key ingredient to the success of platforms such as Netflix, Amazon and Expedia. Product variety has exploded, catering to the heterogeneous tastes of customers. However, this has also increased search costs, making it difficult for customers to find products that interest them. Platforms add value by learning a customer's preferences over time, and leveraging this information to match her with relevant products.

The personalized recommendation problem is typically formulated as an instance of collaborative filtering (Sarwar et al. 2001, Linden et al. 2003). In this setting, the platform observes different customers' past ratings or purchase decisions for random subsets of products. Collaborative filtering

techniques use the feedback across all observed customer-product pairs to infer a low-dimensional model of customer preferences over products. This model is then used to make personalized recommendations over unseen products for any specific customer. While collaborative filtering has found industry-wide success (Breese et al. 1998, Herlocker et al. 2004), it is well-known that it suffers from the "cold start" problem (Schein et al. 2002). In particular, when a new customer enters the platform, no data is available on her preferences over *any* products. Collaborative filtering can only make sensible personalized recommendations for the new customer after she has rated at least $\mathcal{O}(d \log n)$ products, where d is the dimension of the low-dimensional model learned via collaborative filtering and n is the total number of products. Consequently, bandit approaches have been proposed in tandem with collaborative filtering (Bresler et al. 2014, Li et al. 2016, Gopalan et al. 2016) to tackle the cold start problem using a combination of exploration and exploitation. The basic idea behind these algorithms is to offer random products to customers during an exploration phase, learn the customer's low-dimensional preference model, and then exploit this model to make good recommendations.

A key assumption underlying this literature is that customers are patient, and will remain on the platform for the entire (possibly unknown) time horizon T regardless of the goodness of the recommendations that have been made thus far. However, this is a tenuous assumption, particularly when customers have strong outside options (e.g., a Netflix user may abandon the platform for Hulu if they receive a series of bad entertainment recommendations). We demonstrate this effect using customer panel data on a series of ad campaigns from a major commercial airline. Specifically, we find that a customer is far more likely to click on a suggested travel product in the current ad campaign if the previous ad campaign's recommendation was relevant to her. In other words, customers may *disengage* from the platform and ignore new recommendations entirely if past recommendations were irrelevant. In light of this issue, we introduce a new formulation of the bandit product recommendation problem where customers may disengage from the platform depending on the rewards of past recommendations, i.e., the customer's time horizon T on the platform is no longer fixed, but is a function of the platform's actions thus far.

Customer disengagement introduces a significant difficulty to the dynamic learning or bandit literature. We prove lower bounds that show that any algorithm in this setting achieves regret that scales linearly in T (the customer's time horizon on the platform if they are given good recommendations). This hardness result arises because no algorithm can satisfy *every* customer early on when we have limited knowledge of their preferences; thus, no matter what policy we use, at least some customers will disengage from the platform. The best we can hope to accomplish is to keep a large fraction of customers engaged on the platform for the entire time horizon, and to match these customers with their preferred products.

However, classical bandit algorithms perform particularly badly in this setting – we prove that *every* customer disengages from the platform with probability one as T grows large. This is because bandit algorithms *over-explore*: they rely on an early exploration phase where customers are offered random products that are likely to be irrelevant for them. Thus, it is highly probable that the customer receives several bad recommendations during exploration, and disengages from the platform entirely. This exploration is continued for the entire time horizon, T, under the principal of optimism. This is not to say that learning through exploration is a bad strategy. We show that a greedy exploitation-only algorithm also under-performs by either over-exploring through natural exploration, or under-exploring by getting stuck in sub-optimal fixed points. Consequently, the platform misses out on its key value proposition of learning customer preferences and matching them to their preferred products.

Our results demonstrate that one needs to more carefully balance the exploration-exploitation tradeoff in the presence of customer disengagement. We propose a simple modification of classical bandit algorithms by constraining the space of possible product recommendations upfront. We leverage the rich information available from existing customers on the platform to identify a diverse subset of products that are palatable to a large segment of potential customer types; all recommendations made by the platform for new customers are then constrained to be in this set. This approach guarantees that mainstream customers remain on the platform with high probability, and that they are matched to their preferred products over time; we compromise on tail customers, but these customers are unlikely to show up on the platform, and catering recommendations to them endangers the engagement of mainstream customers. We formulate the initial optimization of the product offering as an integer program. We then prove that our proposed algorithm achieves sublinear regret in T for a large fraction of customers, i.e., it succeeds in keeping a large fraction of customers on the platform for the entire time horizon, and matches them with their preferred product. Numerical experiments on synthetic and real data demonstrate that our approach significantly improves both regret and the length of time that a customer is engaged with the platform compared to both classical bandit and greedy algorithms.

1.1. Main Contributions

We highlight our main contributions below:

1. *Empirical evidence of disengagement:* We first present empirical evidence of customer disengagement using a sequence of ad campaigns from a major airline carrier. Our results strongly suggest that customers decide to stay on the platform based on the quality of recommendations.

2. *Disengagement model:* A linear bandit is the classical formulation for learning product recommendations for new customers. Motivated by our empirical results on customer disengagement,

we propose a novel formulation, where the customer's horizon length is endogenously determined by past recommendations, i.e., the customer may exit if given poor recommendations.

3. *Hardness & classical approaches:* We show that no algorithm can achieve sub-linear regret in this setting, i.e., customer disengagement introduces substantial difficulty to the dynamic learning problem. Even worse, we show that classical bandit and greedy algorithms over-explore and fail to keep *any* customer engaged on the platform.

4. *Algorithm:* We propose the Constrained Bandit algorithm, which modifies standard bandit strategies by constraining the product set upfront using a novel integer programming formulation. Unlike classical approaches, the Constrained Bandit provably achieves sublinear regret for a significant fraction of customers.

5. *Numerical experiments:* Extensive numerical experiments on synthetic and real world movie recommendation data (we use the publicly available MovieLens data by Harper and Konstan 2016) demonstrate that the Constrained Bandit significantly improves both regret and the length of time that a customer is engaged with the platform. We find that our approach increases mean customer engagement time on MovieLens by up to 80% over classical bandit and greedy algorithms.

1.2. Related Literature

Personalized decision-making is increasingly a topic of interest, and a central problem is that of learning customer preferences and optimizing the resulting recommendations. However, customer disengagement can introduces a significant difficulty to traditional learning algorithms that have been proposed in the literature.

Personalized Recommendations: The value of personalizing the customer experience has been recognized for a long time (Surprenant and Solomon 1987). We refer the readers to Murthi and Sarkar (2003) for an overview of personalization in operations and revenue management applications. Recently, Besbes et al. (2015), Demirezen and Kumar (2016), and Farias and Li (2017) have proposed novel methods for personalization in online content and product recommendations. We take the widely-used collaborative filtering framework (Sarwar et al. 2001, Su and Khoshgoftaar 2009) as our point of departure. However, all these methods suffer from the cold start problem (Schein et al. 2002). When a new customer enters the platform, no data is available on her preferences over any products, making the problem of personalized recommendations challenging.

Bandits: Consequently, bandit approaches have been proposed in tandem with collaborative filtering (Bresler et al. 2014, Li et al. 2016, Gopalan et al. 2016) to tackle the cold start problem using a combination of exploration and exploitation. The basic idea behind these algorithms is to offer random products to customers during an exploration phase, learn the customer's preferences over products, and then exploit this model to make good recommendations. Relatedly, Lika et al.

(2014) and Wei et al. (2017) use machine learning techniques such as similarity measures and deep neural networks to alleviate the cold start problem. In this paper, we consider the additional challenge of customer disengagement, which introduces a significant difficulty to the dynamic learning or bandit literature. In fact, we show that traditional bandit approaches over-explore, and fail to keep any customer engaged on the platform in the presence of disengagement.

At a high level, our work also relates to the broader bandit literature, where a decision-maker must dynamically collect data to learn and optimize an unknown objective function. For example, many have studied the problem of dynamically pricing products with unknown demand (see, e.g., den Boer and Zwart 2013, Keskin and Zeevi 2014, Qiang and Bayati 2016). Agrawal et al. (2016) analyze the problem of optimal assortment selection with unknown user preferences. Johari et al. (2017) learn to match heterogeneous workers (supply) and jobs (demand) on a platform. Kallus and Udell (2016) use online learning for personalized assortment optimization. These studies rely on optimally balancing the exploration-exploitation tradeoff under bandit feedback. Relatedly, Shah et al. (2018) study bandit learning where the platform's decisions affects the arrival process of new customers; interestingly, they find that classical bandit algorithms can perform poorly due to under-exploration. Closer to our findings, Russo and Van Roy (2018) argue that bandit algorithms can over-explore when an approximately good solution suffices, and propose constraining exploration to actions with sufficiently uncertain rewards. A key assumption underlying this literature is that the time horizon T is fixed and independent of the goodness of the decisions made by the decision-maker. We show that this is a tenuous assumption for recommender systems, since customers may disengage from the platform when offered poor recommendations. Thus, the customer's time horizon T is endogenously determined by the platform's actions, necessitating a novel analysis.

Customer Disengagement: Customer disengagement and its relation to service quality have been extensively studied. For instance, Venetis and Ghauri (2004) use a structural model to establish that service quality contributes to long term customer relationship and retention. Bowden (2009) models the differences in engagement behaviour across new and repeat customers. Sousa and Voss (2012) study the impact of e-service quality on customer behavior in multi-channel services.

Closer to our work, Fitzsimons and Lehmann (2004) use a large-scale experiment on college students to demonstrate that poor recommendations can have a considerably negative impact on customer engagement. We show a similar effect of poor recommendations creating customer disengagement on airline campaign data. It is worth noting that Fitzsimons and Lehmann (2004) studies a a single interaction between users and a recommender, while we study the impact of repeated interactions, which is critical for dynamic learning of a customer's preferences. Relatedly, Tan et al. (2017) empirically find that increasing product variety on Netflix *increases* demand concentration around popular products; this is surprising since one may expect that increasing product variety

would cater to the long tail of customers, enabling more nuanced customer-product matches. However, increasing product variety also increases customer search costs, which may cause customers to cluster around popular products or disengage from the platform entirely. Our proposed algorithm, the Constrained Bandit, makes a similar tradeoff — we constrain our recommendations upfront to a set of popular products that cater to mainstream customers. This approach guarantees that mainstream customers remain engaged with high probability; we compromise on tail customers, but these customers are unlikely to show up, and catering recommendations to them endangers the engagement of mainstream customers.

There are also several papers that study service optimization to improve customer engagement. For example, Davis and Vollmann (1990) develop a framework for relating customer wait times with service quality perception, while Lu et al. (2013) provide empirical evidence of changes in customer purchase behavior due to wait times. Kanoria et al. (2018) model customer disengagement based on the goodwill model of Nerlove and Arrow (1962). In their work, a service provider has two options: a low-cost service level with high likelihood of customer abandonment, or a high-cost service level with low likelihood of customer abandonment. Similarly, Aflaki and Popescu (2013), model the customer disengagement decision as a deterministic known function of service quality. None of these papers study learning in the presence of customer disengagement.

A notable exception is Johari and Schmit (2018), who study the problem of learning a customer's tolerance level in order to send an appropriate number of marketing messages without creating customer disengagement. Here, the decision-maker's objective is to learn the customer's tolerance level, which is a scalar quantity. Similar to our work, the customer's disengagement decision is endogenous to the platform's actions (e.g., the number of marketing messages). However, in our work, we seek to learn a low-dimensional model of the customer's preferences, i.e., a complex mapping of unknown customer-specific latent features to rewards based on product features. The added richness in our action space (product recommendations rather than a scalar quantity) necessitates a different algorithm and analysis. Our work bridges the gap between state-of-the-art machine learning techniques (collaborative filtering and bandits) and the extensive modeling literature on customer disengagement and service quality optimization.

2. Motivation

We use customer panel data from a major commercial airline, obtained as part of client engagement at IBM, to provide evidence for customer disengagement. The airline conducted a sequence of ad campaigns over email to customers that were registered with the airline's loyalty program. Our results suggest that a customer indeed disengages with recommendations if a past recommendation was irrelevant to her. This finding motivates our problem formulation described in the next section.

2.1. Data

The airline conducted 7 large-scale non-targeted ad campaigns over the course of a year. Each campaign involved emailing loyalty customers destination recommendations hand-selected by a marketing team at discounted rates. Importantly, these recommendations were made uniformly across customers regardless of customer-specific preferences.

Our sample consists of 130,510 customers. For each campaign, we observe whether or not the customer clicked on the link provided in the email after viewing the recommendations. We assume that a click signals a positive reaction to the recommendation, while no click could signal either (i) a negative reaction to the recommendation, or (ii) that the customer is already disengaged with the airline campaign and is no longer responding to recommendations.

2.2. Empirical Strategy

Since recommendations were not personalized, we use the heterogeneity in customer preferences to understand customer engagement in the current campaign as a function of the customer-specific quality of recommendations in previous campaigns. To this end, we use the first 5 campaigns in our data to build a score that assesses the relevance of a recommendation to a particular customer. We then evaluate whether the quality of the recommendation in the 6^{th} (previous) campaign affected the customer's response in the 7^{th} (current) campaign after controlling for the quality of the recommendation in the 7^{th} (current) campaign. Our reasoning is as follows: in the absence of customer disengagement, the customer's response to a campaign should depend only on the quality of the current campaign's recommendations; if we instead find that the quality of the previous campaign's recommendations plays an additional negative role in the likelihood of a customer click in the current campaign, then this strongly suggests that customers who previously received bad recommendations have disengaged from the airline campaigns.

We construct a personalized relevance score of recommendations for each customer using click data from the first 5 campaigns. This score is trained using the standard collaborative filtering package available in Python, and achieves an in-sample RMSE of 10%. A version of this score was later implemented in practice by the airline for making personalized recommendations to customers in similar ad campaigns, suggesting that it is an effective metric for evaluating customer-specific recommendation quality.

2.3. Regression Specification

We perform our regression over the 7^{th} (current) campaign's click data. Specifically, we wish to understand if the quality of the recommendation in the 6^{th} (previous) campaign affected the customer's response in the current campaign after controlling for the quality of the current campaign's recommendation. For each customer i, we use the collaborative filtering model to evaluate the

relevance score $prev_i$ of the previous campaign's recommendations and the relevance score $curr_i$ of the current campaign's recommendation. We then perform a simple logistic regression as follows:

$$y_i = f(\beta_0 + \beta_1 \cdot prev_i + \beta_2 \cdot curr_i + \varepsilon_i)\,,$$

where f is the logistic function and y_i is the click outcome for customer i in the current campaign, and ε_i is i.i.d. noise. We fit an intercept term β_0, the effect of the previous campaign's recommendation quality on the customer's click likelihood β_1, and the effect of the current campaign's recommendation quality on the customer's click likelihood β_2. We expect β_2 to be positive since better recommendations in the current campaign should yield higher click likelihood in the current campaign. Our null hypothesis is that $\beta_1 = 0$, and a finding that $\beta_1 < 0$ would suggest that customers disengage from the campaigns if previous recommendations were of poor quality.

2.4. Results

Our regression results are shown in Table 1. As expected, we find that customers are more likely to click if the current campaign's recommendation is relevant to the customer, i.e., $\beta_2 > 0$ (p-value = 0.02). More importantly, we find evidence for customer disengagement since customers are less likely to click in the current campaign if the *previous* campaign's recommendation was not relevant to the customer, i.e., $\beta_2 > 0$ (p-value = 7×10^{-9}). In fact, our point estimates suggest that the disengagement effect dominates the value of the current campaign's recommendation since the coefficient β_1 is roughly three times the coefficient β_2. In other words, it is much more important to have offered a relevant recommendation in the previous campaign (i.e., to keep customers engaged with the campaigns) compared to offering a relevant recommendation in the current campaign to get high click likelihood. These results motivate the problem formulation in the next section explicitly modeling customer disengagement.

Variable	**Point Estimate**	**Standard Error**
(Intercept)	-3.62^{***}	0.02
Relevance Score of Previous Ad Campaign	0.06^{***}	0.01
Relevance Score of Current Ad Campaign	0.02^{*}	0.01

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 1 Regression results from airline ad campaign panel data.

3. Problem Formulation

3.1. Preliminaries

We embed our problem within the popular product recommendation framework of collaborative filtering (Sarwar et al. 2001, Linden et al. 2003). In this setting, the key quantity of interest is a matrix $A \in \mathbb{R}^{m \times n}$, whose entries A_{ij} are numerical values rating the relevance of product j to

customer i. Most of the entries in this matrix are missing since a typical customer has only evaluated a small subset of available products. The key idea behind collaborative filtering is to use a low-rank decomposition

$$A = U^\top V\,,$$

where $U \in \mathbb{R}^{m\times d}, V \in \mathbb{R}^{d\times n}$ for some small value of d. The decomposition can be interpreted as follows: each customer $i \in \{1,...,m\}$ is associated with some low-dimensional vector $U_i \in \mathbb{R}^d$ (row i of the matrix U) that models her preferences; similarly, each product $j \in \{1,...,n\}$ is associated with a low-dimensional vector $V_j \in \mathbb{R}^d$ (given by column j of the matrix V) that models its attributes. Then, the relevance or utility of product j to customer i is simply $U_i^\top V_j$. We refer the reader to Su and Khoshgoftaar (2009) for an extensive review of the collaborative filtering literature. We assume that the platform has a large base of existing customers from whom we have already learned good estimates of the matrices U and V. In particular, all existing customers are associated with known vectors $\{U_i\}_{i=1}^m$, and similarly all products are associated with known vectors $\{V_j\}_{j=1}^n$.

Now, consider a single new customer that arrives to the platform. She forms a new row in A, and all the entries in her row are missing since she is yet to view any products. Like the other customers, she is associated with some vector $U_0 \in \mathbb{R}^d$ that models her preferences, i.e., her expected utility for product $j \in \{1,...,n\}$ is $U_0^\top V_j$. However, U_0 is unknown because we have no data on her product preferences yet. We assume that $U_0 \sim \mathcal{P}$, where $\mathcal{P}$ is a known distribution over new customers' preference vectors; typically, $\mathcal{P}$ is taken to be the empirical distribution of known preference vectors associated with the existing customer base $\{U_1,...,U_m\}$. For ease of exposition and analytical tractability, we will take $\mathcal{P}$ to be a multivariate normal distribution $\mathcal{N}(0,\sigma^2 I_d)$ throughout the rest of the paper.

At each time t, the platform makes a single product recommendation $a_t \in \{V_1,...,V_n\}$, and observes a noisy signal of the customer's utility

$$U_0^\top a_t + \varepsilon_t\,,$$

where ε_t is ξ-subgaussian noise. For instance, platforms often make recommendations through email marketing campaigns (see Figure 1 for example emails from Netflix and Amazon), and observe noisy feedback from the customer based on their subsequent click/view/purchase behavior. We seek to learn U_0 through the customer's feedback from a series of product recommendations in order to eventually offer her the best available product on the platform

$$V_* = \arg\max_{V_j \in \{V_1,...,V_n\}} U_0^\top V_j\,.$$

We impose that $U_0^\top V_* > 0$, i.e., the customer receives positive utility from being matched to her most preferred product on the platform; if this is not the case, then the platform is not appropriate

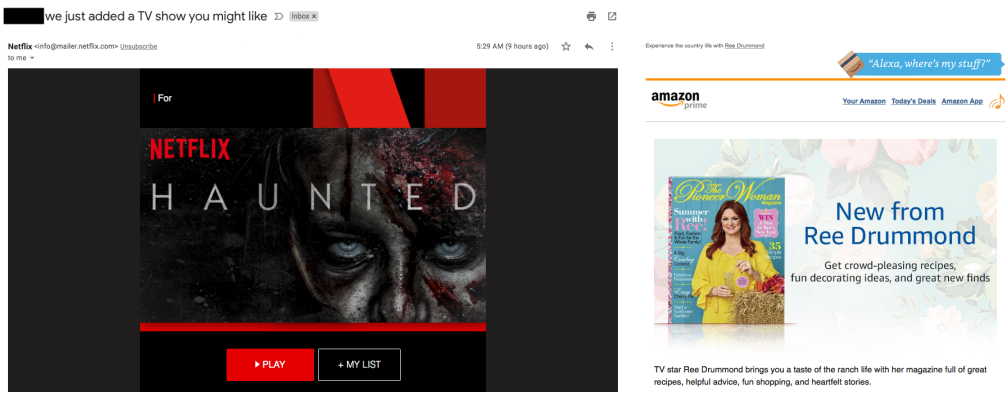

Figure 1 **Examples of personalized recommendations through email marketing campaigns from Netflix (left) and Amazon Prime (right).**

for the customer. We further assume that the product attributes V_i are bounded, i.e., there exists $L > 0$ such that

$$\|V_i\|_2 \leq L \quad \forall i\,.$$

The problem of learning U_0 now reduces to a classical linear bandit (Rusmevichientong and Tsitsiklis 2010), where we seek to learn an unknown parameter U_0 given a discrete action space $\{V_j\}_{j=1}^n$ and stochastic linear rewards. However, as we describe next, our formulation as well as our definition of regret departs from the standard setting by modeling customer disengagement.

3.2. Disengagement Model

Let T be the time horizon for which the customer will stay on the platform if she remains engaged throughout her interaction with the platform. Unfortunately, poor recommendations can cause the customer to disengage from the platform. In particular, at each time t, upon viewing the platform's product recommendation a_t, the customer makes a choice $d_t \in \{0,1\}$ on whether to disengage. The choice $d_t = 1$ signifies that the customer has disengaged and receives zero utility for the remainder of the time horizon T; on the other hand, $d_t = 0$ signifies that the customer has chosen to remain engaged on the platform for the next time period.

There are many ways to model disengagement. For simplicity, we consider the following stylized model: each customer has a tolerance parameter $\rho > 0$ and a disengagement propensity $p \in [0,1]$. Then, the probability that the customer disengages at time t (assuming she has been engaged until now) upon receiving recommendation a_t is:

$$\Pr[d_t = 1 \mid a_t] = \begin{cases} 0 & \text{if } U_0^\top a_t \geq U_0^\top V_* - \rho\,, \\ p & \text{otherwise.} \end{cases}$$

In other words, each customer is willing to tolerate a utility reduction of up to ρ from a recommendation with respect to her utility from her (unknown) optimal product V_*. If the platform makes a

recommendation that results in a utility reduction greater than ρ, the customer will disengage with probability p. Note that we recover the classical linear bandit formulation (with no disengagement) when $p = 0$ or $\rho \to \infty$.

We seek to construct a sequential decision-making policy $\pi = \{a_1, \cdots, a_T\}$ that learns U_0 over time to maximize the customer's utility on the platform. We measure the performance of π by its *cumulative expected regret*, where we modify the standard metric in the analysis of bandit algorithms (Lai and Robbins 1985) to accommodate customer disengagement. In particular, we compare the performance of our policy π against an oracle policy π^* that knows U_0 in advance and always offers the customer her preferred product V_*. At time t, we define the instantaneous expected regret of the policy π for a new customer with realized latent attributes u_0:

$$r_t^\pi(\rho, p, u_0) = \begin{cases} u_0^\top V_* & \text{if } \ d_{t'} = 1 \ \text{ for any } \ t' < t\,, \\ u_0^\top V_* - u_0^\top a_t & \text{otherwise.} \end{cases}$$

This is simply the expected utility difference between the oracle's recommendation and our policy's recommendation, accounting for the fact that the customer receives zero utility for all future recommendations after she disengages. The expectation is taken with respect to ε_t, the ξ-subgaussian noise in realized customer utilities that was defined earlier. The cumulative expected regret for a given customer is then simply

$$\mathcal{R}^\pi(T, \rho, p, u_0) = \sum_{t=1}^{T} r_t^\pi(p, \rho, u_0)\,. \tag{1}$$

Our goal is to find a policy π that minimizes the cumulative expected regret for a new customer whose latent attributes U_0 is a random variable drawn from the distribution $\mathcal{P} = \mathcal{N}(0, \sigma^2 I_d)$. We will show in the next section that no policy can hope to achieve sublinear regret for *all* realizations of U_0; however, we can hope to perform well on likely realizations of U_0, i.e., mainstream customers.

We note that our algorithms and analysis assume that ρ (the tolerance parameter) and p (the disengagement propensity) are known. In practice, these may be unknown parameters that need to be estimated from historical data, or tuned during the learning process. We discuss one possible estimation procedure of these parameters from historical movie recommendation data in our numerical experiments (see Section 6).

To aid the reader, a summary of all variables and their definitions is provided in Table 2 in Appendix A.

4. Classical Approaches

We now prove lower bounds that demonstrate (i) no policy can perform well on every customer in this setting, and (ii) bandit algorithms and greedy Bayesian updating can fail for all customers.

4.1. Preliminaries

We restrict ourselves to the family of non-anticipating policies $\Pi : \pi = \{\pi_t\}$ that form a sequence of random functions π_t that depend only on observations collected until time t. In particular, if we let $H_t = (a_1, Y_1, a_2, Y_2, ...a_{t-1}, Y_{t-1})$ denote the vectorized history of product recommendations and corresponding utility realizations and $\mathcal{F}_t$ denote the σ-field generated by H_t, then π_{t+1} is $\mathcal{F}_t$ measurable. All policies assume full knowledge of the tolerance parameter ρ, the disengagement propensity p, and the distribution of latent customer attributes $\mathcal{P}$.

Next, we define a general class of bandit learning algorithms that achieve sublinear regret in the standard setting with no disengagement.

DEFINITION 1. A bandit algorithm $\pi \in \Pi$ is consistent if for all u_0, there exists $\nu \in [0, 1)$ and $\mathcal{R}(T, \rho, p = 0, u_0) = \mathcal{O}(T^\nu)$. This is equivalent to the following condition:

$$\lim_{T\to\infty} \sup \frac{\log\left(R(T, \rho, p = 0, u_0)\right)}{\log(T)} = \nu\,,$$

where the supremum is taken over all feasible realizations of the unknown customer feature vector u_0. As discussed before, when $p = 0$, our regret definition reduces to the classical bandit regret with no disengagement. The above definition implies that a policy π is consistent if its rate of cumulative regret is sublinear in T. This class (Π^C) includes the well-studied UCB (e.g., Auer 2002, Abbasi-Yadkori et al. 2011) Thompson Sampling, (e.g., Agrawal and Goyal 2013, Russo and Van Roy 2014), and other bandit algorithms. Our definition of consistency is inspired by Lattimore and Szepesvari (2016), but encompasses a larger class of policies. We will show that any algorithm in Π^C fails to perform well in the presence of disengagement.

Notation: For any vector $V \in \mathbb{R}^d$ and positive semidefinite matrix $X \in \mathbb{R}^{d\times d}$, $\|V\|_X$ refers to the operator norm of V with respect to matrix X given by $\sqrt{V^\top XV}$. Similarly, for any set S, $S\backslash i$ for some $i \in S$ refers to the set S without element i. I_d refers to the $d \times d$ identity matrix for some $d \in \mathbb{Z}$. For any series of scalars (vectors), $Y_1, ...Y_t$, $Y_{1:t}$ refers to the column vector of the scalars (vectors) Y_1,..,Y_t. Next, we define the set $\mathcal{S}(u_0, \rho)$ of products that are tolerable to the customer, i.e., recommending any product from this (unknown) set will not cause disengagement:

DEFINITION 2. Let $\mathcal{S}(u_0, \rho)$ be the set of products, amongst all products, that satisfy the tolerance threshold for the customer with latent attribute vector, u_0. More specifically, when $p > 0$,

$$\mathcal{S}(u_0, \rho) := \{i : u_0^\top V_i \;\geq\; u_0^\top V_* - \rho, \forall i = 1, .., n\}\,. \tag{2}$$

Note that in the classical bandit setting, this set contains all products, $|\mathcal{S}(u_0, \rho)| = n$. When $\mathcal{S}(u_0, \rho)$ is large, exploration is less costly, but as the customer tolerance threshold ρ decreases, $|\mathcal{S}(u_0, \rho)|$ decreases as well.

Finally, we consider the following simplified latent product features to enable a tractable analysis.

EXAMPLE 1 (SIMPLE SETTING). We assume that there are d total products in $\mathbb{R}^d$, and the latent product features $V_i = e_i$, the i^{th} basis vector. We also take $p > 0$, i.e., customers may disengage.

4.2. Lower bounds

We first show an impossibility result that no non-anticipating policy can obtain sublinear regret over *all* customers. We consider the worst-case regret of any non-anticipating policy over all feasible customer tolerance parameters ρ.

THEOREM 1 **(Hardness Result)**. *Under the assumptions of Example 1, any non anticipating policy $\pi \in \Pi$ achieves regret that scales linearly with T:*

$$\inf_{\pi \in \Pi} \sup_{\rho > 0} \mathbb{E}_{u_0 \sim \mathcal{P}} \left[\mathcal{R}^{\pi}(T, \rho, p, u_0) \right] = C \cdot T = \mathcal{O}(T) \,,$$

where $C \in \mathbb{R}$ is a constant independent of T but dependent on other problem parameters.

Proof: See Appendix B. □

Theorem 1 shows that the expected worst case regret is linear in T. In other words, regardless of the policy chosen, there exists a subset of customers (with positive measure under $\mathcal{P}$) who incur linear regret in the presence of disengagement. The proof relies on showing that there is always a positive probability that the customer (i) will not be offered her preferred product in the first time step, and consequently, (ii) for sufficiently small ρ, will disengage from the platform immediately. Thus, in expectation, any non-anticipating policy is bound to incur linear regret.

Theorem 1 shows that product recommendation with customer disengagement requires making a trade-off over the types of customers that we seek to engage. No policy can keep *all* the users engaged without knowing the user's preference apriori. Nevertheless, since Theorem 1 only characterizes the worst case *expected* regret, this poor performance can be caused by a very small fraction of customers. Hence, another approach could be to ensure that at least a large fraction of customers (mainstream customers) are engaged, while potentially sacrificing the engagement of customers with niche preferences (tail customers).

In Theorem 2, we show that consistent bandit learning algorithms fail to achieve engagement even for mainstream customers throughout the time horizon. Thus, in contrast to showing that the worst case *expected* regret is linear (Theorem 1), we show that the worst case regret is linear for *any* customer realization u_0.

THEOREM 2 **(Failure of Bandits)**. *Let u_0 be any realization of the latent user attributes from $\mathcal{P}$. Under the assumptions of Example 1, any consistent bandit algorithm $\pi \in \Pi^C$ achieves regret that scales linearly with T for this customer as $T \to \infty$. That is,*

$$\inf_{\pi \in \Pi^C} \sup_{\rho > 0} \mathcal{R}^{\pi}(T, \rho, p, u_0) = C_1 \cdot T = \mathcal{O}(T) \,,$$

where $C_1 \in \mathbb{R}$ is a constant independent of T but dependent on other problem parameters.

Proof: See Appendix B. □

Theorem 2 shows that the worst case regret of consistent bandit policies is linear for *every* customer realization (including mainstream customers). We note that this result is worse than what we may have hoped for given the earlier hardness result (Theorem 1), since the linearity of regret applies to all customers rather than a subset of customers. The proof of Theorem 2 considers the case when the size of the set of tolerable products $|\mathcal{S}(u_0, \rho)| < d$, which occurs for sufficiently small ρ. Clearly, exploring outside this set can lead to customer disengagement. However, since $|\mathcal{S}(u_0, \rho)| < d$, this set of products cannot span the space $\mathbb{R}^d$, implying that one cannot recover the true customer latent attributes u_0 without sampling products outside of the set. On the other hand, consistent bandit algorithms require convergence to u_0, i.e., they will sample outside the set $\mathcal{S}(u_0, \rho)$ infinitely many times (as $T \to \infty$) at a rate that depends on their corresponding regret bound. Yet, it is clear to see that offering infinitely many recommendations outside the customer's set of tolerable products $\mathcal{S}(u_0, \rho)$ will eventually lead to customer disengagement (when $p > 0$) with probability 1. This result highlights the tension between avoiding incomplete learning (which requires exploring products outside the tolerable set) and avoiding customer disengagement (which requires restricting our recommendations to the tolerable set). Thus, we see that the design of bandit learning strategies fundamentally relies on the assumption that the time horizon T is exogeneous, making exploration inexpensive. State-of-the-art techniques such as UCB and Thompson Sampling perform particularly poorly by over-exploring in the presence of customer disengagement.

Recent literature has highlighted the success of greedy policies in bandit problems where exploration may be costly (see, e.g., Bastani et al. 2017). One may expect that the natural exploration afforded by greedy policies may enable better performance in settings where exploration can lead to customer disengagement. Therefore, we now shift our focus to Greedy Bayesian Updating policy (Algorithm 1) below. We use a Bayesian policy since we wish to make full use of the known prior $\mathcal{P}$ over latent customer attributes. Unfortunately, we find that, similar to consistent bandit algorithms, the greedy policy also incurs worst-case linear regret for *every* customer. Furthermore, the greedy policy can perform poorly even when there is no disengagement.

The greedy Bayesian updating policy begins by recommends the most commonly preferred product based on the $\mathcal{P}$. Then, in every subsequent time step, it observes the customer response, updates its posterior on the customer's latent attributes using Bayesian linear regression, and then offers the most commonly preferred product based on the updated posterior. The form of the resulting estimator $\hat{u}_t$ of the customer's latent attributes is similar to the well-known ridge regression estimator with regularization parameter $\frac{\xi^2}{\sigma^2 t}$, where we regularize towards the mean of the prior $\mathcal{P}$ over latent customer attributes (which we have normalized to 0 here).

Algorithm 1 Greedy Bayesian Updating (GBU)

Initialize and recommend a randomly selected product.
for $t \in [T]$ **do**
  Observe customer utility, $Y_t = u_0^\top a_t + \varepsilon_t$.
  Update customer feature estimate, $\hat{u}_{t+1} = \left(a_{1:t}^\top a_{1:t} + \frac{\xi^2}{\sigma^2} I\right)^{-1} (a_{1:t}^\top Y_{1:t})$.
  Recommend product $a_{t+1} = \arg\max_{i=1\ldots,n} \hat{u}_{t+1}^\top V_i$.
end for

In Theorem 3, we show that the greedy policy also fails to achieve engagement even for mainstream customers throughout the time horizon. In essence, the *free exploration* induced by greedy policies (see, e.g., Bastani et al. 2017, Qiang and Bayati 2016) is in theory as problematic as the optimistic exploration by bandit algorithms. Furthermore, Theorem 4 shows that even when exploration is not costly (there is no disengagement), the greedy policy can get stuck at suboptimal fixed points, and fail to produce a good match.

THEOREM 3 **(Failure of Greedy)**. *Let u_0 be any realization of the latent user attributes from $\mathcal{P}$. Under the assumptions of Example 1, the GBU policy achieves regret that scales linearly with T for this customer as $T \to \infty$. That is,*

$$\sup_{\rho > 0} \mathcal{R}^{GBU}(T, \rho, p, u_0) = C_2 \cdot T = \mathcal{O}(T)\,,$$

where $C_2 \in \mathbb{R}$ is a constant independent of T but dependent on other problem parameters.

Proof: See Appendix B. □

Similar to our result for consistent bandit algorithms in Theorem 2, Theorem 3 shows that the worst case regret of the greedy policy is linear for *every* customer realization (including mainstream customers). While intuition may suggest that greedy algorithms avoid over-exploration, they still involve natural exploration due to the noise in customer feedback, which may cause the algorithm to over-explore and choose irrelevant products. Although Theorems 2 and 3 are similar, it is worth noting that over-exploration is *empirically* much less likely with the greedy policy than with a consistent bandit algorithm that is designed to explore. This difference is exemplified in our numerical experiments in §6; however, we will see that one is still better off (both theoretically and empirically) constraining exploration by restricting the product set upfront.

The proof of Theorem 3 has two cases: tail and mainstream customers. For tail customers (this set is determined by the choice of ρ), the first offered product (the most commonly preferred product across customers given the distribution $\mathcal{P}$) may not be tolerable, and so they disengage immediately with some probability p, yielding linear expected regret for these customers. Note that this is true for any algorithm, including the Constrained Bandit. The more interesting case is that of mainstream customers, who *do* find the first offered product tolerable. In this case, since customer

feedback is noisy, the greedy policy may subsequently erroneously switch to a product outside of the tolerable set, which again results in immediate customer disengagement with probability p. Note that this effect is exactly the natural exploration that allows the greedy policy to sometimes yield rate-optimal convergence in classical contextual bandits (Bastani et al. 2017). Putting these two cases together, we find that the greedy policy achieves linear regret for every customer.

It is also worth considering the performance of the greedy policy when there is no disengagement and exploration is not costly. In Theorem 4, we show that the greedy policy may under-explore and fail to converge in the other extreme, i.e., when there is no customer disengagement. Note that, unlike the previous results, this result is under the case of $p=0$ (otherwise, the setting of Example 1 applies).

THEOREM 4 **(Failure of Greedy without Disengagement)**. *Let $\rho \to \infty$ or $p=0$, i.e., there is no customer disengagement. The GBU policy achieves regret that scales linearly with T. That is,*

$$\mathbb{E}_{u_0 \sim \mathcal{P}} \left[\mathcal{R}^{GBU}(T, \rho, p=0, u_0) \right] = C_3 \cdot T = \mathcal{O}(T)\,,$$

where $C_3 \in \mathbb{R}$ is a constant independent of T but dependent on other problem parameters.

Proof: See Appendix B. □

Theorem 4 shows that the greedy policy fails with some probability even in the classical bandit learning setting when there is no customer disengagement. The proof follows from considering the subset of customers for whom the most commonly preferred product is *not* their preferred product. We show that within this subset, the greedy policy continues recommending this suboptimal product for the remaining time horizon T with positive probability. This illustrates that a greedy policy can get "stuck" on a suboptimal product due to incomplete learning (see, e.g., Keskin and Zeevi 2014) even when customers never disengage. Thus, we see that the greedy policy can also fail due to *under-exploration*. In contrast, a consistent bandit policy is always guaranteed to converge to the preferred product when there is no disengagement; the Constrained Bandit will trivially achieve the same guarantee since we will not restrict the product set when there is no disengagement.

These results illustrate that there is a need to constrain exploration in the presence of customer disengagement; however, naively adopting a greedy policy does not achieve this goal. This is because, intuitively, the greedy policy constrains the *rate* of exploration rather than the *size* of exploration. The proof of Theorem 2 clearly demonstrates that the key issue is to constrain exploration to be within the set of tolerable products $\mathcal{S}(u_0, \rho)$. The challenge is that this set is unknown since the customer's latent attibutes u_0 are unknown. However, our prior $\mathcal{P}$ gives us reasonable knowledge of which products lie in $\mathcal{S}(u_0, \rho)$ for mainstream customers. In the next section, we will leverage this knowledge to restrict the product set upfront in the Constrained Bandit. As we saw

from Theorem 1, we may as well restrict our focus to serving the subset of mainstream customers, since we cannot hope to do well for all customers.

5. Constrained Bandit Algorithm

We have so far established that both classical bandit algorithms and the greedy algorithm may fail to perform well on *every* customer. We now propose a two-step procedure, where we play a bandit strategy after constraining our action space to a restricted set of products that are carefully chosen using an integer program. In §5.3, we will prove that this simple modification guarantees good performance on a significant fraction of customers.

5.1. Intuition

As shown in Theorem 2, classical bandit algorithms fail because of over-exploration. Bandit algorithms rely on an early exploration phase where customers are offered random products; the feedback from these products is then used to infer the customer's low-dimensional preference model, in order to inform future (relevant) recommendations during the exploitation phase. However, in the presence of customer disengagement, the algorithm doesn't get to reap the benefits of exploitation since the customer likely disengages from the platform during the exploration phase after receiving several irrelevant recommendations. This is not to say that learning through exploration is a bad strategy. Theorem 3 shows that greedy exploitation-only algorithm also under-perform by under-exploring, and getting stuck in sub-optimal fixed points. This can be harmful since the platform misses out on its key value proposition of learning customer preferences and matching them to their preferred products.

These results suggest that a platform can only succeed by avoiding poor early recommendations. Since we don't know the customer's preferences, this is impossible to do in general; however, our key insight is that a probabilistic approach is still feasible. In particular, the platform has knowledge of the distribution of customer preferences $\mathcal{P}$ from past customers, and can transfer this knowledge to avoid products that do not meet the tolerance threshold of most customers. We formulate this product selection problem as an integer program, which ensures that any recommendations within the optimal restricted set are acceptable to most customers. After selecting an optimal restricted set of products, we follow a classical bandit approach (e.g., linear UCB by Abbasi-Yadkori et al. 2011). Under this approach, if our new customer is a mainstream customer, she is unlikely to disengage from the platform even during the exploration phase, and will be matched to her preferred product. However, if the new customer is a tail customer, her preferred product may not be available in our restricted set, causing her to disengage. This result is shown formally in Theorem 6 in the next section. Thus, we compromise performance on tail customers to achieve good performance on

mainstream customers. Theorem 1 shows that such a tradeoff is necessary, since it is impossible to guarantee good performance on *every* customer.

We introduce a set diameter parameter γ in our integer program formulation. This parameter can be used to tune the size of the restricted product set based on our prior $\mathcal{P}$ over customer preferences. Larger values of γ increase the risk of customer disengagement by introducing greater variability in product relevance, but also increase the likelihood that the customer's preferred product lies in the set. On the other hand, smaller values of γ decrease the risk of customer disengagement *if* the customer's preferred product is in the restricted set, but there is a higher chance that the customer's preferred product is not in the set. Thus, appropriately choosing this parameter is a key ingredient of our proposed algorithm. We discuss how to choose γ at the end of §5.3.

5.2. Constrained Exploration

We seek to find a restricted set of products that cater to a large fraction of customers (which is measured with respect to the distribution $\mathcal{P}$ over customer attributes), but are not too "far" from each other (to limit exploration). Before we describe the problem, we introduce notation that captures the likelihood of a product being relevant for the new customer:

DEFINITION 3. $\mathcal{C}_i(\rho)$ is the probability of product i satisfying the new customer's tolerance level:

$$\mathcal{C}_i(\rho) = \mathbb{P}_{u_0 \sim \mathcal{P}}(i \in \mathcal{S}(u_0, \rho)),$$

where $\mathcal{S}(u_0, \rho)$ is given by Definition 2.

Recall that $\mathcal{S}(u_0, \rho)$ is the set of tolerable products for a customer with latent attributes u_0. Given that u_0 is unknown, $\mathcal{C}_i(\rho)$ captures the probability that product i is relevant to the customer with respect to the distribution $\mathcal{P}$ over random customer preferences. In the presence of disengagement, we seek to explore over products that are likely to satisfy the new customer's tolerance level. For example, mainstream products may be tolerable for a large probability mass of customers (with respect to $\mathcal{P}$) while niche products may only be tolerable for tail customers. Thus, $\mathcal{C}_i(\rho)$ translates our prior on customer latent attributes to a likelihood of tolerance over the space of products. Computing $\mathcal{C}_i(\rho)$ using Monte Carlo simulation is straightforward: we generate random customer latent attributes according to $\mathcal{P}$, and count the fraction of customers for which product i was within the customer's tolerance threshold of ρ from the customer's preferred product V_*.

As discussed earlier, a larger product set increases the likelihood that the new customer's preferred product is in the set, but it also increases the likelihood of disengagement due to poor recommendations during the exploration phase. However, the key metric here is not the number of products in the set, but rather the similarity of the products in the set. In other words, we

wish to restrict product diversity in the set to ensure that all products are tolerable to mainstream customers. Thus, we define

$$D_{ij} = \|V_i - V_j\|_2 \,,$$

the Euclidean distance between the (known) features of products i and j, i.e., the similarity between two products. We seek to find a subset of products such that the distance between any pair of products is bounded by the set diameter γ. Let $\phi_{ij}(\gamma)$ be an indicator function that determines whether $D_{ij} \leq \gamma$. Hence,

$$\phi_{ij}(\gamma) = \begin{cases} 1 & \text{if } D_{ij} \leq \gamma \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

Note that γ and ρ are related. When the customer tolerance ρ is large, we will choose larger values of the set diameter γ and vice-versa. We specify how to choose γ at the end of §5.3.

The objective is to select a set of products, which together have a high likelihood of containing the customer's preferred match under the distribution over customer preferences $\mathcal{P}$ (i.e., high $\mathcal{C}_i(\rho)$), with the constraint that no two products are too dissimilar from each other (i.e., pairwise distance greater than γ). We propose solving the following product selection integer program:

$$\mathbf{OP}(\gamma) = \max_{\mathbf{x},\mathbf{z}} \ \sum_{i=1}^{n} C_i(\rho) x_i \tag{3a}$$

$$\text{s.t. } z_{ij} \leq x_i, \quad i = 1, \ldots, n, \tag{3b}$$

$$z_{ij} \leq x_j, \quad j = 1, \ldots, n, \tag{3c}$$

$$z_{ij} \geq x_i + x_j - 1, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, \tag{3d}$$

$$z_{ij} \leq \phi_{ij}(\gamma), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, \tag{3e}$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, n. \tag{3f}$$

The decision variables in the above problem are $\{x_i\}_{i=1}^n$ and $\{z_{i,j}\}_{i,j=1}^n$. In particular, x_i in $\mathbf{OP}(\gamma)$ defines whether product i is included in the restricted set, and $z_{i,j}$ is an indicator variable for whether both products i and j are included in the restricted set. Constraints (3b) – (3e) ensure that only products that are "close" to each other are selected.

Solving $\mathbf{OP}(\gamma)$ results in a set of products (products for which the corresponding x_i is 1) that maximizes the likelihood of satisfying the new customer's tolerance level, while ensuring that every pair is within γ distance from each other.

Algorithm 2 presents the Constrained Bandit (CB) algorithm, where the second phase follows the popular linear UCB algorithm (Abbasi-Yadkori et al. 2011). There are two input parameters: λ (the standard regularization parameter employed in the linear bandit literature, see, e.g., Abbasi-Yadkori et al. 2011) and γ (the set diameter). We discuss the selection of γ and the corresponding

Algorithm 2 Constrained Bandit(λ,γ)

Step 1: Constrained Exploration:
Solve $\mathbf{OP}(\gamma)$ to get Ξ, the constrained set of products to explore over. Let a_1 be a randomly selected product to recommend in Ξ.
Step 2: Bandit Learning:
for $t \in [T]$ **do**
Observe customer utility, $Y_t = u_0^\top a_t + \varepsilon_t$.
Let $\hat{u}_t = (a_{1:t}^\top a_{1:t} + \lambda I)^{-1} a_{1:t} Y_{1:t}$, and,

$$\mathcal{Q}_t = \left\{ u \in \mathbb{R}^d : \|\hat{u}_t - u\|_{\bar{X}_t} \leq \left(\xi \sqrt{d \log\left(\frac{1+tL^2}{\delta}\right)} + \sqrt{\lambda}\frac{\rho}{\gamma} \right) \right\}.$$

Let $(u_{opt}, a_t) = \arg\max_{\{i \in \Xi, u \in \mathcal{Q}_t\}} u^\top V_i$.
Recommend product a_t at time t if the customer is still engaged. Stop if the customer disengages from the platform.
end for

tradeoffs in the next subsection and in Appendix D. As discussed earlier, we employ a two-step procedure. In the first step, the action space is restricted to the product set given by $\mathbf{OP}(\gamma)$. This step ensures that subsequent exploration is unlikely to cause a significant fraction of customers to disengage. Then, a standard bandit algorithm is used to learn the customer's preference model and match her with her preferred product through repeated interactions. The main idea remains simple: in the presence of customer disengagement, the platform should be cautious while exploring. Since we are uncertain about the customer's preferences, we optimize exploration for mainstream customers who are more likely to visit the platform.

5.3. Theoretical Guarantee

We now show that the Constrained Bandit performs well and incurs regret that scales sublinearly in T over a fraction of customers. We begin by defining $L_{t,\rho,p}$, an indicator variable that captures whether the customer is still engaged at time t:

Definition 4. Let,

$$L_{t,\rho,p} = \begin{cases} 1 & \text{Customer engaged until time t}, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathbb{1}\{L_{T,\rho,p} = 1\} = \Pi_{t=1}^{T} \mathbb{1}\{d_t = 0\},$$

where we recall that d_t is the disengagement decision of the customer at time t. To show our result, we first show that as $T \to \infty$, $L_{T,\rho,p} = 1$ for some customers, i.e., they remain engaged. Next, we show that most engaged customers are eventually matched to their preferred product.

Theorem 5 shows that the worst-case regret of the Constrained Bandit scales sublinearly in T for a positive fraction of customers. In particular, regardless of the customer tolerance parameter

ρ, we can match some subset of customers to their preferred products. Note that this is in stark contrast with both bandit and greedy algorithms (Theorems 2 and 3).

THEOREM 5 **(Matching Upper Bound for Constrained Bandit)**. *Let u_0 be any realization of the latent user attributes from $\mathcal{P}$. Under the assumptions of Example 1, the Constrained Bandit with set diameter $\gamma = 1/\sqrt{2}$ achieves zero regret with positive probability. In particular, there exists $\mathcal{W}_{\lambda,\gamma=\frac{1}{\sqrt{2}}}$, a set of realizations of customer latent attributes with positive measure under $\mathcal{P}$, i.e.,*

$$\mathbb{P}\left(\mathcal{W}_{\lambda,\gamma=\frac{1}{\sqrt{2}}}\right) > 0\,,$$

such that, for all $u_0 \in \mathcal{W}_{\lambda,\gamma=\frac{1}{\sqrt{2}}}$, the worst-case regret of the Constrained Bandit algorithm is

$$\sup_{\rho>0} \mathcal{R}^{\mathrm{CB}\left(\lambda,\gamma=\frac{1}{\sqrt{2}}\right)}(T,\rho,p,u_0) = 0\,.$$

Note that this result holds for any value of ρ, i.e., customers can be arbitrarily intolerant of products that are not their preferred product V_*. Thus, the only way to make progress is to immediately recommend their preferred product. This can trivially be done by restricting our product set to a single product, which at the very least caters to *some* customers. This is exactly what we do in Theorem 5: the choice of $\gamma = 1/\sqrt{2}$ and the product space given in Example 1 ensures that only a single product will be in our restricted set Ξ. By construction of $\mathbf{OP}(\gamma)$, this will be the most popular preferred product. $\mathcal{W}$ denotes the subset of customers for whom this product is optimal, and this set has positive measure under $\mathcal{P}$ by construction since we have a discrete number of products. Note that these customers are immediately matched to their preferred product, so it immediately follows that we incur zero regret on this subset of customers.

Theorem 5 shows that there is nontrivial value in restricting the product set upfront, which cannot be obtained through either bandit or greedy algorithms. However, it considers the degenerate case of constraining exploration to only a single product, which is clearly too restrictive in practice, especially when customers are relatively tolerant (i.e., ρ is not too small). Thus, it does not provide useful insight into how much the product set should be constrained as a function of the customer's tolerance parameter. To answer this question, we move away from the setting described in Example 1 and consider a fluid approximation of the product space. Since the nature of $\mathbf{OP}(\gamma)$ is complex, letting the product space be continuous $V = [-1,1]^d$ will help us cleanly demonstrate the key tradeoff in constraining exploration: a larger product set has a higher probability of containing customers' preferred products, but also a higher risk of disengagement. Furthermore, for algebraic simplicity, we shift the mean of the prior over the customer's latent attributes, so $\mathcal{P} = \mathcal{N}(\bar{u}, \frac{\sigma^2}{d} I_d)$, where $\|\bar{u}\|_2 = 1$. This ensures that our problem is not symmetric, which again helps us analytically characterize the solution of $\mathbf{OP}(\gamma)$.

Theorem 6 shows that the Constrained Bandit algorithm can achieve sublinear regret for a fraction of customers under this albeit stylized setting. More importantly, it yields insights into how we might choose the set diameter γ as a function of the customer's tolerance parameter ρ. In §6, we demonstrate the strong empirical performance of our algorithm on real data.

THEOREM 6 **(Guarantee for Constrained Bandit Algorithm)**. *Let $\mathcal{P} = \mathcal{N}(\bar{u}, \frac{\sigma^2}{d} I_d)$. Also consider a continuous product space $V = [-1,1]^d$. There exists a set $\mathcal{W}$ of latent customer attribute realizations with positive probability under $\mathcal{P}$, i.e.,*

$$\mathbb{P}(\mathcal{W}) \geq w \;=\; \left(1 - 2d \exp\left(-\frac{1 - \sqrt{\left(1 - \frac{\gamma^2}{4}\right)}}{\sigma}\right)\right)\left(1 - 2d \exp\left(-\left(\frac{\frac{\rho}{\gamma} - \sum_{i=1}^{i=d} \bar{u}_i}{\sigma}\right)^2\right)\right),$$

such that for all $u_0 \in \mathcal{W}$ the cumulative regret of the Constrained Bandit is

$$\begin{aligned}\mathcal{R}^{CB(\lambda,\rho)}(T, \rho, p, u_0) &\leq 5\sqrt{Td \log\left(\lambda + \frac{TL}{d}\right)}\left(\sqrt{\lambda}\frac{\rho}{\gamma} + \xi\sqrt{\log(T) + d\log\left(1 + \frac{TL}{\lambda d}\right)}\right) \\ &= \tilde{\mathcal{O}}\left(\sqrt{T}\right).\end{aligned}$$

Proof: See Appendix C. □

This result explicitly characterizes the fraction of customers that we successfully serve as a function of the customer tolerance parameter ρ and the set diameter γ. Thus, given a value of ρ, we can choose the set diameter γ to optimize the probability w of this set.

The proof of Theorem 6 follows in three steps. First, we lower bound the probability that the constrained exploration set Ξ contains the preferred product for a new customer whose attributes are drawn from $\mathcal{P}$. Next, conditioned on the previous event, we lower bound the probability that the customer remains engaged for the entire time horizon T when recommendations are made from the restricted product set Ξ. Lastly, conditioned on the previous event, we can apply standard self-normalized martingale techniques (Abbasi-Yadkori et al. 2011) to bound the regret of the Constrained Bandit algorithm for the customer subset $\mathcal{W}$.

Again, as in Theorem 5, we see that there can be significant value in restricting the product set upfront that cannot be achieved by classical bandit or greedy approaches. We further see that the choice of the set diameter γ is an important consideration to ensure that the new customer is engaged and matched to her preferred product with as high a likelihood as possible. As discussed earlier, larger values of γ increase the risk of customer disengagement by introducing greater variability in product relevance, but also increase the likelihood that the customer's preferred product lies in the set. On the other hand, smaller values of γ decrease the risk of customer disengagement *if* the customer's preferred product is in the restricted set, but there is a higher

chance that the customer's preferred product is not in the set. In other words, we wish to choose γ to maximize w. While there is no closed form expression for the optimal γ, we propose the following approximately optimal choice based on a Taylor series approximation (see details in Appendix D):

$$\gamma^* \in \left\{\gamma : \rho = \frac{\sqrt{\sigma}\gamma^2}{2\left(4-\gamma^2\right)^{1/4}} \text{ and } \gamma > 0\right\}.$$

Numerical experiments demonstrate that this approximate value of γ is typically within 1% of the value of γ that maximizes the expression for w given in Theorem 6; the resulting values of w are also very close (see Appendix D). This expression yields some interesting comparative statics: we should choose a smaller set diameter γ when customers are less tolerant (ρ is small) and customer feedback is noisy (σ is large). In practice, we can tune the set diameter through cross-validation.

6. Numerical Experiments

We now compare the empirical performance of the Constrained Bandit with the state-of-the-art Thompson sampling (which is widely considered to empirically outperform other bandit algorithms, see, e.g., Chapelle and Li 2011, Russo and Van Roy 2014) and a greedy Bayesian updating policy. We present two sets of empirical results evaluating our algorithm on both synthetic data (§6.1), and on real movie recommendation data (§6.2).

Benchmarks: We compare our algorithm with (i) linear Thompson Sampling (Russo and Van Roy 2014) and (ii) the greedy Bayesian updating (Algorithm 1).

Constrained Thompson Sampling (CTS): To ensure a fair comparison, we consider a Thompson Sampling version of the Constrained Bandit algorithm (see Algorithm 3 below). Recall that our approach allows for any bandit strategy after obtaining a restricted product set based on our (algorithm-independent) integer program $\mathbf{OP}(\gamma)$. We use the same implementation of linear Thompson sampling (Russo and Van Roy 2014) as our benchmark in the second step. Thus, any improvements in performance can be attributed to restricting the product set.

Algorithm 3 Constrained Thompson Sampling (λ,γ)

Step 1: Constrained Exploration:
Solve $\mathbf{OP}(\gamma)$ to get the constrained set of products to explore over, $S_{constrained}$. Let $\hat{u}_1 = \bar{u}$.
Step 2: Bandit Learning:
for $t \in [T]$ **do**
  Sample $u(t)$ from distribution $\mathcal{N}(\hat{u}_t, \sigma^2 I_d)$.
  Recommend $a_t = \arg\max_{\{i \in S_{constrained}\}} u(t)^\top V_i$ if the customer is still engaged.
  Observe customer utility, $Y_t = U_0^\top a_t + \varepsilon_t$, and update $\hat{u}_t = (V_{a_1:a_t}^\top V_{a_1:a_t} + \lambda I)^{-1} V_{a_1:a_t} Y_{1:t}$
  Stop if the customer disengages from the platform.
end for

6.1. Synthetic Data

We generate synthetic data and study the performance of all three algorithms as we increase the customer's disengagement propensity $p \in [0,1]$. A low value of p implies that customer disengagement is not a salient concern, and thus, one would expect Thompson sampling to perform well in this regime. On the other hand, a high value of p implies that customers are extremely intolerant of poor recommendations, and thus, all algorithms may fare poorly. We find that Constrained Thompson Sampling performs comparably to vanilla Thompson Sampling when p is low, and offers sizeable gains over both benchmarks when p is medium or large.

Data generation: We consider the standard collaborative filtering problem (described earlier) with 10 products. Recall that collaborative filtering fits a low rank model of latent customer preferences and product attributes; we take this rank[1] to be 2. We generate product features from a multivariate normal distribution with mean $[1,5]^\top \in \mathbb{R}^2$ and variance $0.3 \cdot I_2 \in \mathbb{R}^{2\times 2}$, where we recall that I_d is the $d \times d$ identity matrix. Similarly, latent user attributes are generated from a multivariate normal with with mean $[2,2]^\top \in \mathbb{R}^2$ and variance $2 \cdot I_2 \in \mathbb{R}^{2\times 2}$. These values ensure that, with high probability for every customer, there exists a product on the platform that generates positive utility. Note that the product features are known to the algorithms, but the latent user attributes are unknown. Finally, we take our noise $\epsilon \sim \mathcal{N}(0,5)$, the customer tolerance ρ to be generated from a truncated $\mathcal{N}(0,1)$ distribution, and the total horizon length $T = 1000$. All algorithms are provided with the distribution of customer latent attributes, the distribution of the customer tolerance ρ, and the horizon length T. They are not provided with the noise variance, which needs to be estimated over time. Finally, we consider several values of the disengagement propensity, i.e., $p \in \{1\%, 10\%, 50\%, 100\%\}$, to capture the value of restricting the product set with varying levels of customer disengagement.

Engagement Time: We use average customer engagement time (i.e., the average time that a customer remains engaged with the platform, up to time T) as our metric for measuring algorithmic performance. As we have seen in earlier sections, customer engagement is necessary to achieve low cumulative regret. Furthermore, it is a more relevant metric from a managerial perspective since higher engagement is directly related with customer retention and loyalty, as well as the potential for future high quality/revenue customer-product matches.

Results: Figure 2 shows the customer engagement time averaged over 1000 randomly generated users (along with the 95% confidence intervals) for all three algorithms as we vary the disengagement propensity p from 1% to 100%. As expected, when $p = 1\%$ (i.e., customer disengagement is relatively insignificant), TS performs well, and CTS performs comparably. However, as noted in

[1] We choose a small rank based on empirical experiments showing that collaborating filtering models perform better in practice with small rank (Chen and Chi 2018). Our results remain qualitatively similar with higher rank values.

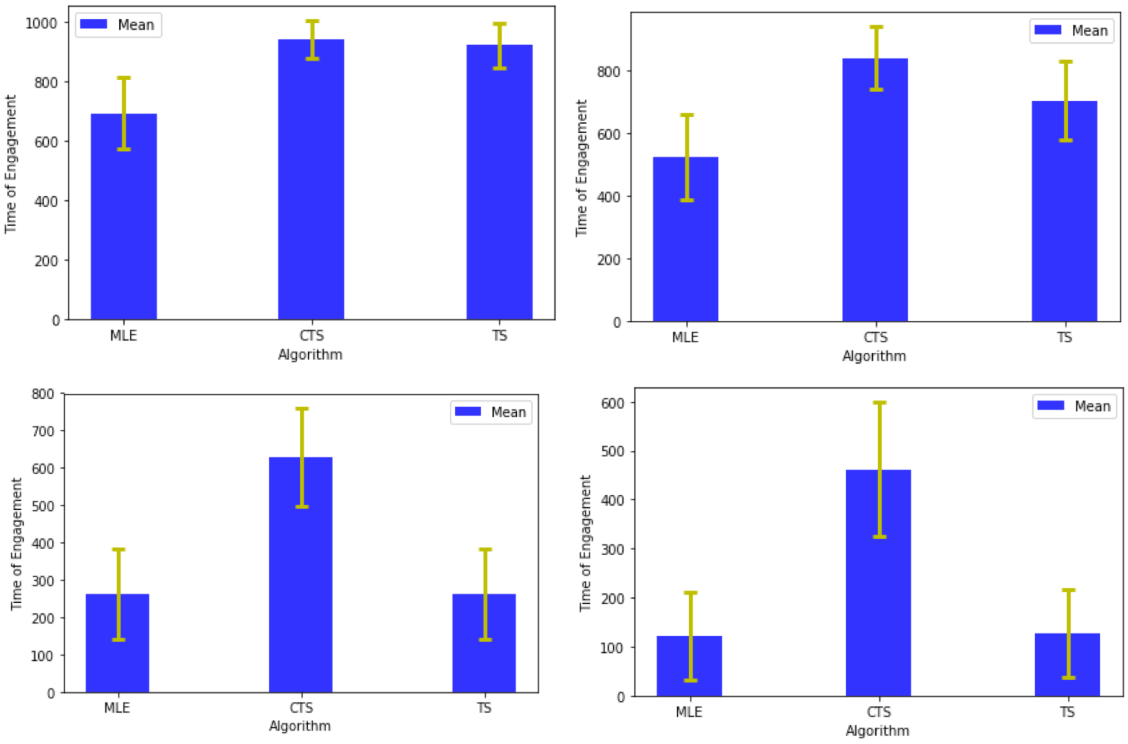

Figure 2 Time of engagement and 95% confidence intervals averaged over 1000 randomly generated customers for disengagement propensity p values of 1% (top left), 10% (top right), 50% (bottom left), and 100% (bottom right).

Theorem 3, greedy Bayesian updating is likely to converge to a suboptimal product outside of the customer's relevance set, and continues to recommend this product until the customer eventually disengages. As we increase p, all algorithms achieve worse engagement, since customers become considerably more likely to leave the platform. As expected, we also see that CTS starts to significantly outperform the other two benchmark algorithms as p increases. For instance, the mean engagement time of CTS improves over the engagement time of the benchmark algorithms by a factor of 2.2 when $p = 50\%$ and by a factor or 4.4 when $p = 100\%$. Thus, we see that restricting the product set is critical when customer disengagement is a salient feature on the platform.

A recent report by Smith (2018) notes that an average worker receives as many as 121 emails on average per day. Furthermore, the average click rate for retail recommendation emails is as low as 2.5%. These number suggest that customer disengagement is becoming increasingly salient, and we argue that constraining exploration on these platforms to quickly match as many customers as possible to a tolerable product is a key consideration in recommender system design.

6.2. Case Study: Movie Recommendations

We now compare CTS to the same benchmarks on MovieLens, a publicly available movie recommendations data collected by GroupLens Research. This dataset is widely used in the academic

community as a benchmark for recommendation and collaborative filtering algorithms (Harper and Konstan 2016). Importantly, we no longer have access to the problem parameters (e.g., ρ) and must estimate them; we discuss simple heuristics for estimating these parameters.

6.2.1. Data Description & Parameter Estimation The MovieLens dataset contains over 20 million user ratings based on personalized recommendations of 27,000 movies to 138,000 users. We use a random sample (provided by MovieLens) of 100,000 ratings from 671 users over 9,066 movies. Ratings are made on a scale of 1 to 5, and are accompanied by a time stamp for when the user submitted the rating. The average movie rating is 3.65.

The first step in our analysis is identifying likely disengaged customers in our data. We will argue that the number of user ratings is a proxy for disengagement. In Figure 3, we plot the histogram of the number of ratings per user. Users provide an average of 149 ratings, and a median of 71 ratings. Clearly, there is high variability and skew in the number of ratings that users provide. We

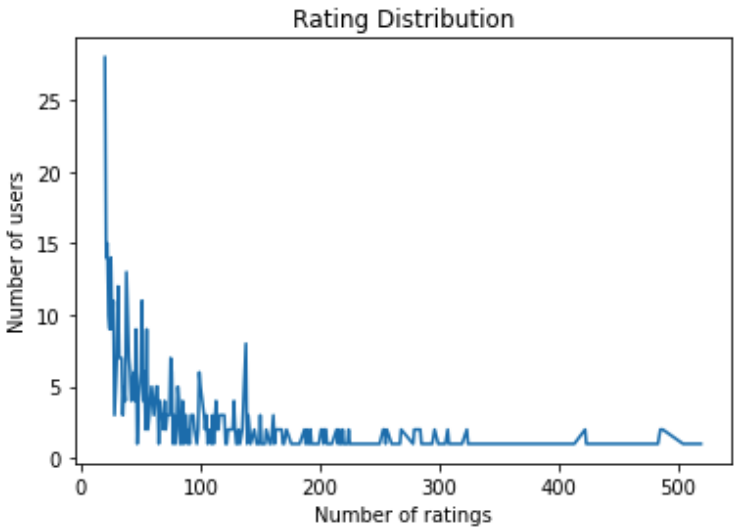

Figure 3 Histogram of user ratings in MovieLens data.

argue that there are two primary reasons why a customer may stop providing ratings: (i) satiation and (ii) disengagement. Satiation occurs when the user has exhausted the platform's offerings that are relevant to her, while disengagement occurs when the user is relatively new to the platform and does not find sufficiently relevant recommendations to justify engaging with the platform. Thus, satiation applies primarily to users who have provided many ratings (right tail of Figure 3), while disengagement applies primarily to users who have provided very few ratings (left tail of Figure 3).

Accordingly, we consider the subset of users who provided fewer that 27 ratings (bottom 15% of users) as *disengaged* users. We hypothesize that these users provided a low number of ratings

because they received recommendations that did not meet their tolerance threshold. This hypothesis is supported by the ratings. In particular, the average rating of disengaged users is 3.56 (standard error of 0.10) while the average rating of the remaining (engaged) users is 3.67 (standard error of 0.04). A one-way ANOVA test (Welch 1951) yields a F-statistic of 29.23 and a p-value of 10^{-8}, showing that the difference is statistically significant and that disengaged users dislike their recommendations more than engaged users. This finding relates to our results in §2, i.e., disengagement is related to the customer-specific quality of recommendations made by the platform.

Estimating latent user and movie features: We need to estimate the latent product features $\{V_i\}_{i=1}^n$ as well as the distribution $\mathcal{P}$ over latent user attributes from historical data. Thus, we use low rank matrix factorization (Ekstrand et al. 2011) on the ratings data (we find that a rank of 5 yields a good fit) to derive $\{U_i\}_{i=1}^m$ and $\{V_i\}_{i=1}^n$. We fit a normal distribution $\mathcal{P}$ to the latent user attributes $\{U_i\}_{i=1}^m$, and use this to generate new users; we use the latent product features as-is.

Estimating the tolerance parameter ρ: Recall that ρ is the maximum utility reduction (with respect to the utility of the unknown optimal product V_*) that a customer is willing to tolerate before disengaging with probability p. In our theory, we have so far assumed that there is a single known value of ρ for all customers. However, in practice, it is likely that ρ may be a random value that is sampled from a distribution (e.g., there may be natural variability in tolerance among customers), and further, the distribution of ρ may be different for different customer types (e.g., tail customer types may be more tolerant of poor recommendations since they are used to having higher search costs for niche products). Thus, we estimate the distribution of ρ as a function of the user's latent attributes u_0 using maximum likelihood estimation, and sample different realizations for different incoming customers on the platform. We detail the process of this estimation next.

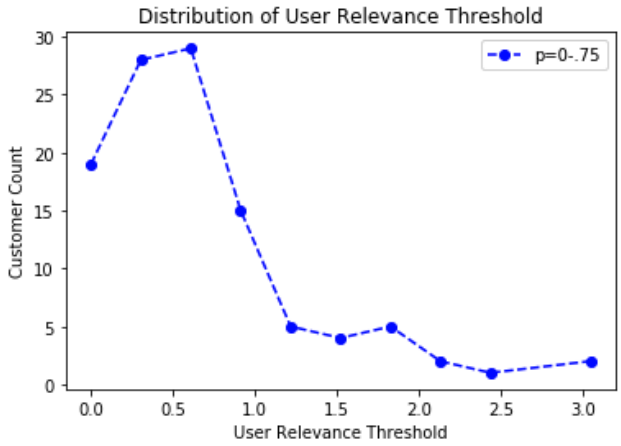

Figure 4 Empirical distribution of ρ, the customer-specific tolerance parameter, across all disengaged users for a fixed customer disengagement propensity $p = .75$. This distribution is robust to any choice of $p \in (0, .75]$.

In order to estimate ρ for a user, we consider the time series of ratings provided by a single user with latent attributes u_0 in our historical data. Clearly, disengagement occurred when the user provided the last rating to the platform, and this decision was driven by both the user's disengagement propensity p, and tolerance parameter ρ. For a given p and ρ, let t^{leave} denote the last rating of the user, and $a_1, ...a_{t_{leave}}$ be the recommendations made to the user until time t_{leave}. Then, the likelihood function of the observation sequence is:

$$\mathcal{L}(p,\rho) = p(1-p)^{\left(t^{leave} - \sum_{i=1}^{(t^{leave}-1)} \mathbb{1}\{a_i \in \mathcal{S}(u_0,\rho)\}\right)},$$

where we recall that $\mathcal{S}(u_0,\rho)$ defines the set of products that the user considers tolerable. Since u_0 and V_i are known apriori (estimated from the low rank model), $\mathcal{S}(u_0,\rho)$ is also known apriori for any given value of ρ. Hence, for any given value of p, we can estimate the most likely user-specific tolerance parameter ρ using the maximum likelihood estimator of $\mathcal{L}(p,\rho)$. In Figure 4, we plot the overall estimated empirical distribution of ρ for our subset of disengaged users. We see that more than 88% of disengaged users have an estimated tolerance parameter of less than 1.2, i.e., they consider disengagement if the recommendation is more than 1 star away from what they would rate their preferred movie. As we may expect, very few disengaged users have a high estimated value of ρ, suggesting that they have high expectations on the quality of recommendations.

One caveat of our estimation strategy is that we are unable to identify both p and ρ simultaneously; instead, we estimate the user-specific distribution of ρ and perform our simulations for varying values of the disengagement propensity p. Empirically, we find that our estimation of ρ is robust to different values of p, i.e., for any value of $p \in (0, .75]$, we observe that our estimated distribution of ρ distribution does not change. Thus, we believe that this strategy is sound.

6.2.2. Results Similar to §6.1, we compare Constrained Thompson Sampling against our two benchmarks (Thompson Sampling and greedy Bayesian updating) based on average customer engagement time. We use a random sample of 200 products, and take our horizon length $T = 100$.

Figure 5 shows the customer engagement time averaged over 1000 randomly generated users (along with the 95% confidence intervals) for all three algorithms as we vary the disengagement propensity p from 1% to 100%. Again, we see similar trends as we saw in our numerical experiments on synthetic data (§6.1). When $p = 1\%$ (i.e., customer disengagement is relatively insignificant), all algorithms perform well, and CTS performs comparably. As we increase p, all algorithms achieve worse engagement, since customers become considerably more likely to leave the platform. As expected, we also see that CTS starts to significantly outperform the other two benchmark algorithms as p increases. For instance, the mean engagement time of CTS improves over the engagement time of the benchmark algorithms by a factor of 1.26 when $p = 10\%$, by a factor of 1.66 when

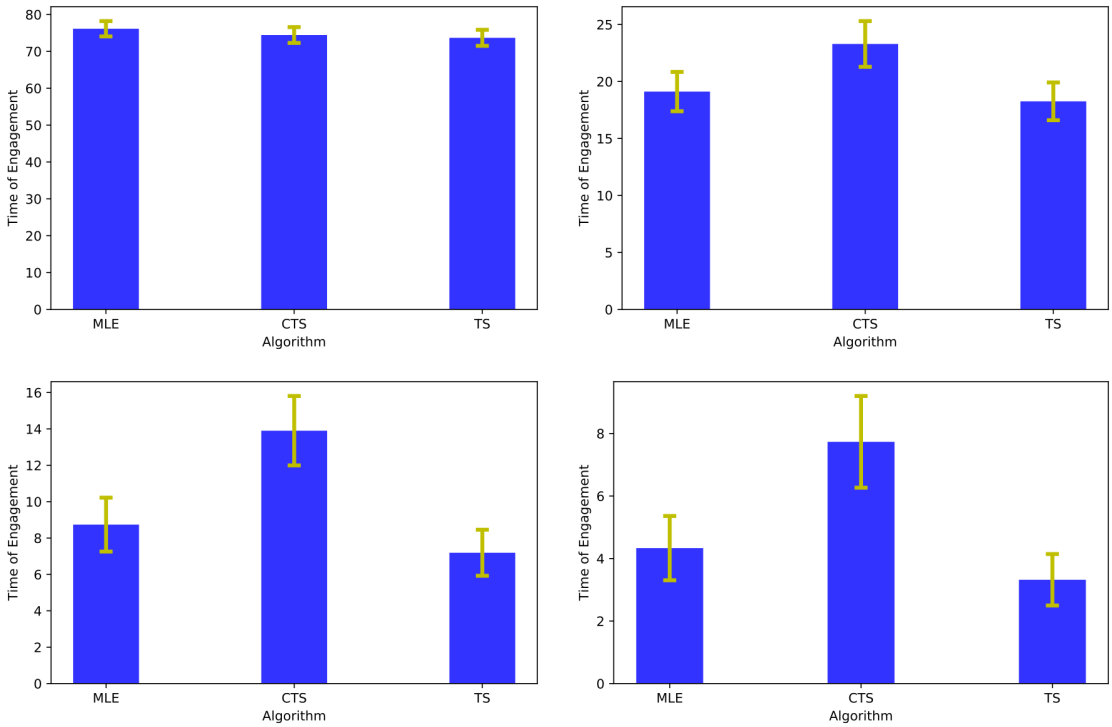

Figure 5 Time of engagement and 95% confidence intervals on MovieLens data averaged over 1000 randomly generated customers for disengagement propensity p values of 1% (top left), 10% (top right), 50% (bottom left), and 100% (bottom right).

$p = 50\%$ and by a factor or 1.8 when $p = 100\%$. Thus, our main finding remains similar on real movie recommendation data: restricting the product set is critical when customer disengagement is a salient feature on the platform.

7. Conclusions

We consider the problem of sequential product recommendation when customer preferences are unknown. First, using a sequence of ad campaigns from a major airline carrier, we present empirical evidence suggesting that customer disengagement plays an important role in the success of recommender systems. In particular, customers decide to stay on the platform based on the quality of recommendations. To the best of our knowledge, this issue has not been studied in the framework of collaborative filtering, a widely-used machine learning technique. We formulate this problem as a linear bandit, with the notable difference that the customer's horizon length is a function of past recommendations. Our formulation bridges two disparate literatures on bandit learning in recommender systems, and customer disengagement modeling.

We then prove that this problem is fundamentally hard, i.e., no algorithm can keep all customers engaged. Thus, we shift our focus to keeping a large number of customers (i.e., mainstream customers) engaged, at the expense of tail customers with niche preferences. Our results highlight

a necessary tradeoff with clear managerial implications for platforms that seek to make personalized recommendations. Unfortunately, we find that classical bandit learning algorithms as well as a simple greedy Bayesian updating strategy perform poorly, and can fail to keep any customer engaged. To solve this problem, we propose modifying bandit learning strategies by constraining the action space upfront using an integer program. We prove that this simple modification allows our algorithm to perform well (i.e., achieve sublinear regret) for a significant fraction of customers. Furthermore, we perform extensive numerical experiments on real movie recommendations data that demonstrate the value of restricting the product set upfront. In particular, we find that our algorithm can improve customer engagement with the platform by up to 80% in the presence of significant customer disengagement.

References

Abbasi-Yadkori, Yasin, Dávid Pál, Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *NIPS*. 2312–2320.

Aflaki, Sam, Ioana Popescu. 2013. Managing retention in service relationships. *Management Science* **60**(2) 415–433.

Agrawal, Shipra, Vashist Avadhanula, Vineet Goyal, Assaf Zeevi. 2016. A near-optimal exploration-exploitation approach for assortment selection. *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 599–600.

Agrawal, Shipra, Navin Goyal. 2013. Further optimal regret bounds for thompson sampling. *Artificial Intelligence and Statistics*. 99–107.

Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**(Nov) 397–422.

Bastani, Hamsa, Mohsen Bayati, Khashayar Khosravi. 2017. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011* .

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2015. Optimization in online content recommendation services: Beyond click-through rates. *Manufacturing & Service Operations Management* **18**(1) 15–33.

Bowden, Jana Lay-Hwa. 2009. The process of customer engagement: A conceptual framework. *Journal of Marketing Theory and Practice* **17**(1) 63–74.

Breese, John S, David Heckerman, Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *UAI*. Morgan Kaufmann Publishers Inc., 43–52.

Bresler, Guy, George H Chen, Devavrat Shah. 2014. A latent source model for online collaborative filtering. *NIPS*. 3347–3355.

Chapelle, Olivier, Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*. 2249–2257.

Chen, Yudong, Yuejie Chi. 2018. Harnessing structures in big data via guaranteed low-rank matrix estimation. *arXiv preprint arXiv:1802.08397* .

Davis, Mark M, Thomas E Vollmann. 1990. A framework for relating waiting time and customer satisfaction in a service operation. *Journal of Services Marketing* **4**(1) 61–69.

Demirezen, Emre M, Subodha Kumar. 2016. Optimization of recommender systems based on inventory. *Production and Operations Management* **25**(4) 593–608.

den Boer, Arnoud V, Bert Zwart. 2013. Simultaneously learning and optimizing using controlled variance pricing. *Management science* **60**(3) 770–783.

Ekstrand, Michael D, John T Riedl, Joseph A Konstan, et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction* **4**(2) 81–173.

Farias, Vivek F, Andrew A Li. 2017. Learning preferences with side information .

Fitzsimons, Gavan J, Donald R Lehmann. 2004. Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science* **23**(1) 82–94.

Gopalan, Aditya, Odalric-Ambrym Maillard, Mohammadi Zaki. 2016. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508* .

Harper, F Maxwell, Joseph A Konstan. 2016. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* **5**(4) 19.

Herlocker, Jonathan L, Joseph A Konstan, Loren G Terveen, John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *TOIS* **22**(1) 5–53.

Johari, Ramesh, Vijay Kamble, Yash Kanoria. 2017. Matching while learning. *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 119–119.

Johari, Ramesh, Sven Schmit. 2018. Learning with abandonment. *arXiv preprint arXiv:1802.08718* .

Kallus, Nathan, Madeleine Udell. 2016. Dynamic assortment personalization in high dimensions. *arXiv preprint arXiv:1610.05604* .

Kanoria, Yash, Ilan Lobel, Jiaqi Lu. 2018. Managing customer churn via service mode control .

Keskin, N Bora, Assaf Zeevi. 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* **62**(5) 1142–1167.

Lai, Tze Leung, Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6**(1) 4–22.

Lattimore, Tor, Csaba Szepesvari. 2016. The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491* .

Li, Shuai, Alexandros Karatzoglou, Claudio Gentile. 2016. Collaborative filtering bandits. *SIGIR*. ACM, 539–548.

Lika, Blerina, Kostas Kolomvatsos, Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert Systems with Applications* **41**(4) 2065–2073.

Linden, Greg, Brent Smith, Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* (1) 76–80.

Lu, Yina, Andrés Musalem, Marcelo Olivares, Ariel Schilkrut. 2013. Measuring the effect of queues on customer purchases. *Management Science* **59**(8) 1743–1763.

Murthi, BPS, Sumit Sarkar. 2003. The role of the management sciences in research on personalization. *Management Science* **49**(10) 1344–1362.

Nerlove, Marc, Kenneth J Arrow. 1962. Optimal advertising policy under dynamic conditions. *Economica* 129–142.

Qiang, Sheng, Mohsen Bayati. 2016. Dynamic pricing with demand covariates .

Rusmevichientong, Paat, John N Tsitsiklis. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* **35**(2) 395–411.

Russo, Daniel, Benjamin Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research* **39**(4) 1221–1243.

Russo, Daniel, Benjamin Van Roy. 2018. Satisficing in time-sensitive bandit learning. *arXiv preprint arXiv:1803.02855* .

Sarwar, Badrul, George Karypis, Joseph Konstan, John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. *WWW*. ACM, 285–295.

Schein, Andrew I, Alexandrin Popescul, Lyle H Ungar, David M Pennock. 2002. Methods and metrics for cold-start recommendations. *SIGIR*. ACM, 253–260.

Shah, Virag, Jose Blanchet, Ramesh Johari. 2018. Bandit learning with positive externalities .

Smith, Craig. 2018. 90 interesting email statistics and facts. URL `https://expandedramblings.com/index.php/email-statistics/`.

Sousa, Rui, Chris Voss. 2012. The impacts of e-service quality on customer behaviour in multi-channel e-services. *Total Quality Management & Business Excellence* **23**(7-8) 789–806.

Su, Xiaoyuan, Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* **2009**.

Surprenant, Carol F, Michael R Solomon. 1987. Predictability and personalization in the service encounter. *the Journal of Marketing* 86–96.

Tan, Tom Fangyun, Serguei Netessine, Lorin Hitt. 2017. Is tom cruise threatened? an empirical study of the impact of product variety on demand concentration. *Information Systems Research* **28**(3) 643–660.

Venetis, Karin A, Pervez N Ghauri. 2004. Service quality and customer retention: building long-term relationships. *European Journal of marketing* **38**(11/12) 1577–1598.

Wei, Jian, Jianhua He, Kai Chen, Yi Zhou, Zuoyin Tang. 2017. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications* **69** 29–39.

Welch, Bernard Lewis. 1951. On the comparison of several mean values: an alternative approach. *Biometrika* **38**(3/4) 330–336.

Appendix

A. Summary of Notation

Variables	***Description***
T	Total time horizon
ρ	Customer specific tolerance threshold
p	Customer specific leaving propensity
$\bar{u}$	prior mean on latent user attributes
σ^2	prior variance on user attributes
ξ	sub Gaussian noise parameter
ν	Sub linear rate of regret for the consistent policy
d_t	Customer's leaving decision at time t
U_0	Random vector denoting customer feature vector
u_0	Realization of U_0
$\mathcal{S}(\rho, u_0)$	User specific set of "relevant" products
Δ_V	Utility gap between the optimal and the sub optimal product
a_t	Recommendation made at time t
Y_t	User utility response at time t
V_*	Optimal product
λ	L2 regularization parameter
γ	set diameter for the integer program
Ξ	Constrain product exploration set resulting from solving $\mathbf{OP}(\gamma)$

Table 2 Summary of notation used in the paper.

B. Lower Bounds for Classical Approaches

Proof of Theorem 1: In order to show the linearity of regret of any policy, we start by considering the simple case when $d = 2$. Recall that, $u_0 \in \mathbb{R}^2$ such that $u_0 \sim \mathcal{N}(0, \sigma^2 I_2)$. Furthermore, by assumption $V_1 = [1, 0]$ and $V_2 = [0, 1]$ are the product attributes and p, the leaving propensity, is some positive constant. Clearly, Product 1 is optimal when $u_{0_1} > u_{0_2}$ and vice versa. Furthermore, if product 1 (product 2) is shown to a customer who prefers product 2 (product 1), then the customer disengages with probability p as long as the gap in utility is more than ρ. Hence, consider the following events:

$$\mathcal{E}_1 = \{u_{0_1} < u_{0_2} - \rho\}, \text{and } \mathcal{E}_2 = \{u_{0_2} < u_{0_1} - \rho\}.$$

Then over $\mathcal{E}_1$, recommending product 1 leads to customer disengagement with probability p and over $\mathcal{E}_2$, recommending product 2 leads to customer disengagement with probability p. Next, we will characterize the probability of events $\mathcal{E}_1$ and $\mathcal{E}_2$.

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(u_{0_1} < u_{0_2} - \rho) &= \mathbb{P}\left(\frac{u_{0_1} - u_{0_2}}{\sqrt{2}\sigma} < \frac{-\rho}{\sqrt{2}\sigma}\right) \\ &= \mathbb{P}\left(Z < \frac{-\rho}{\sqrt{2}\sigma}\right) = C,\end{aligned}$$

such that $C \in (0, 1)$. Similarly,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2) = \mathbb{P}(u_{0_2} < u_{0_1} - \rho) &= \mathbb{P}\left(\frac{u_{0_2} - u_{0_1}}{\sqrt{2}\sigma} < \frac{-\rho}{\sqrt{2}\sigma}\right) \\ &= \mathbb{P}\left(Z < \frac{-\rho}{\sqrt{2}\sigma}\right) = C.\end{aligned}$$

Any policy π has two options at time 1: either to recommend product 1 or to recommend product 2. We will show that regret is linear in either of the two cases. First consider the case when $a_1 = 1$ and note that

$$\begin{aligned}\mathbb{E}_{u_0\sim\mathcal{P}}\left[\mathcal{R}^\pi(T,\rho,p,u_0)\right] = \mathbb{E}_{U_0\sim\mathcal{P}}\left[\sum_{t=1}^{T} r_t(\rho,p,u_0)\right] &\geq \sum_{t=1}^{T} r_t(\rho,p,u_0\in\mathcal{E}_1).\mathbb{P}\left(\mathcal{E}_1\right)\\ &\geq \sum_{t=1}^{T} r_t(\rho,p,u_0\in\mathcal{E}_1)\cdot\mathbb{P}\left(\mathcal{E}_1\right)\cdot\mathbb{1}(d_1=1)\cdot\mathbb{P}(d_1=1)\\ &\geq T\cdot\mathbb{P}\left(\mathcal{E}_1\right)\cdot p\\ &= CpT\,.\end{aligned}$$

One can similarly show that when $a_1 = 2$,

$$\mathbb{E}_{u_0\sim\mathcal{P}}\left[\mathcal{R}^\pi(T,\rho,p,u_0)\right] \geq CpT\,.$$

Hence,

$$\inf_{\pi\in\Pi}\sup_{\rho>0}\mathbb{E}_{U_0\sim\mathcal{P}}\left[\mathcal{R}^\pi(T,\rho,p,U_0)\right] = C\cdot p\cdot T = \mathcal{O}(T)\,,$$

The proof follows similarly for any $d > 2$ and we skip the details here for the sake of brevity. □

Before we prove Theorem 2, we prove an important Lemma that relates the confidence width of the mean reward of product V ($\|V\|^2_{X_t^{-1}}$) and shows that this width shrinks at a rate faster than the confidence width of the estimation of the gap between reward from V and the optimal product (Δ_V).

LEMMA 1. *Let π be a consistent policy and let $a_1,..,a_t$ be actions taken under policy π. Let $u_0 \in R^d$ be a realization of the random user vector, $U_0 \sim \mathcal{P}$, such that there is a unique optimal product, V_* amongst the set of feasible products. Then $\forall\ V \in \{V_1,....V_n\}/\ V_*$*

$$\limsup_{t\to\infty}\log(t)\|V\|^2_{X_t^{-1}} \leq \frac{\Delta_V^2}{2(1-\nu)}\,,$$

where $\Delta_V = u_0^\top V_* - u_0^\top V$ and $X_t = \mathbb{E}\left[\sum_{l=1}^{T} a_l a_l'\right]$.

Proof: We will prove this result in two steps. In Step 1, we will show that

$$\limsup_{t\to\infty}\log(t)\|V - V_*\|^2_{X_t^{-1}} \leq \frac{\Delta_V^2}{2(1-\nu)}\,.$$

In Step 2, we will connect this result to the matrix norm on the features of V which will prove the result.

The proof strategy is similar to that of Theorem 2 in Lattimore and Szepesvari (2016) and follows from two Lemmas that are provided in Appendix E (Lemma 4 and Lemma 5) for the sake of completeness.

Step 1: First note that for any realization of u_0, Δ_v is finite and positive for all sub optimal arms since there is a unique optimal product (V_* for u_0). Now consider any suboptimal arm. Then for any event $A \in \Omega_t$ (measurable sequence of recommendations and utility realizations until time t), and probability measures $\mathbb{P}$ and $\mathbb{P}'$ on Ω_t for a fixed bandit policy, Lemma 4 shows that

$$\mathrm{KL}(\mathbb{P},\mathbb{P}') \geq \log\left(\frac{1}{2\mathbb{P}(A)+\mathbb{P}'(A^c)}\right)\,.$$

Combining the above with Lemma 5 shows that

$$\frac{1}{2}\|u_0 - u_0'\|^2_{X_t} = \mathrm{KL}(\mathbb{P},\mathbb{P}') \geq \log\left(\frac{1}{2\mathbb{P}(A)+\mathbb{P}'(A^c)}\right)\,.$$

where u_0 and u_0' are two different user latent attribute . Now assume that, u_0 and u_0', are chosen such that V_* is not the optimal arm for u_0'. That is, $\exists V$ such that $(V - V_*)^\top u_0' > 0$. Now let $A = \{T_{i*}(t) \leq \frac{t}{2}\}$ where i^* is the index of product V_*. Note that $T_i(t)$ is the total number of times arm i is pulled until time t. Then, A is the event that the optimal product is pulled very few times. By construction, for all consistent policies, such an event will have a very low probability. More precisely, if we let

$$u_0' = u_0 + \frac{H\,(V - V_*)}{\|V - V_*\|_H^2}\,(\Delta_V + \epsilon)\,,$$

for some positive definite matrix H, then it follows that the optimality gap between V and V_* is ϵ. Now, if we consider the expected instantaneous regret for u_0 until time t, then we have that

$$\mathcal{R}(t, u_0) = \sum_{i=1}^{n} \Delta_{V_i} \mathbb{E}\left[T_i(t)\right] \geq \Delta_{min} \mathbb{E}\left[t - T_{i^*}(t)\right] \geq \Delta_{min} \mathbb{E}\left[\mathbb{1}\left\{T_{i^*}(t) \leq \frac{t}{2}\right\}\frac{t}{2}\right] = \frac{t\Delta_{min}}{2}\mathbb{P}\left(T_{i^*}(t) \leq \frac{t}{2}\right),$$

where Δ_{min} is the minimum suboptimality gap. Note that we suppress the dependence of ρ and p for cleaner exposition. Similarly, writing the regret characterization for u_0', we have that

$$\mathcal{R}(t, u_0') = \sum_{i=1}^{n} \Delta_{V_i}^{'} \mathbb{E}'\left[T_i(t)\right] \geq \Delta_{i*} \mathbb{E}'\left[T_{i^*}(t)\right] \geq \frac{t\epsilon}{2}\mathbb{P}'\left(T_{i^*}(t) > \frac{t}{2}\right).$$

where Δ_V', $\mathbb{E}'$ and $\mathbb{P}'$ are the optimality gap, the expectation operator induced by the probability measure $\mathbb{P}'$ over u_0'. We then have that

$$\frac{\mathcal{R}(t, u_0) + \mathcal{R}(t, u_0')}{\epsilon t} \geq \mathbb{P}(A) + \mathbb{P}(A')\,,$$

where we have assumed that ϵ is sufficiently small. Next, considering u_0', we have that

$$\begin{aligned}\frac{1}{2}\|u_0 - u_0'\|_{X_t}^2 = \frac{(\Delta_V + \epsilon)\,\|V - V_*\|_{HX_tH}^2}{2\|V - V_*\|_H^4} &\geq \log\left(\frac{1}{2\mathbb{P}(A) + \mathbb{P}'(A^c)}\right) \\ &\geq \log\left(\frac{\epsilon t}{2\left(\mathcal{R}(t, u_0) + \mathcal{R}(t, u_0')\right)}\right).\end{aligned}$$

Hence, multiplying $\log(t)$ on both sides above, we have that

$$\frac{1}{2\log(t)}\frac{(\Delta_V + \epsilon)\,\|V - V_*\|_{HX_tH}^2}{\|V - V_*\|_H^4} \geq 1 + \frac{\log\left(\frac{\epsilon}{2}\right)}{\log(t)} - \log\left(\mathcal{R}(t, u_0) + \mathcal{R}(t, u_0')\right)\frac{1}{\log(t)}\,.$$

But recall that for any consistent policy,

$$\lim_{t\to\infty} \sup \frac{\log(\mathcal{R}(t, u_0))}{\log(T)} = \nu\,.$$

for some $0 \leq \nu < 1$. Hence, for any positive semidefinite H such that $\|V - V_*\|_H > 0$,

$$\lim_{t\to\infty} \inf \frac{1}{2\log(t)}\frac{(\Delta_V + \epsilon)\,\|V - V_*\|_{HX_tH}^2}{\|V - V_*\|_H^4} = \lim_{t\to\infty} \inf \frac{\|V - V_*\|_{X_t^{-1}}^2}{2\log(t)}\frac{(\Delta_V + \epsilon)\,\|V - V_*\|_{HX_tH}^2}{\|V - V_*\|_{X_t^{-1}}^2\|V - V_*\|_H^4} \geq 1 - \nu\,.$$

Next, consider a subsequence $\{X_{t_k}\}_{k=1}^{\infty}$ such that

$$\tilde{c} = \lim_{t\to\infty} \sup \log(t)\|V - V_*\|_{X_t^{-1}}^2 = \lim_{k\to\infty} \log(t_k)\|V - V_*\|_{X_{t_k}^{-1}}^2\,.$$

Then,

$$\liminf_{t\to\infty} \frac{\|V-V_*\|^2_{X_t^{-1}}}{\log(t)} \frac{\|V-V_*\|^2_{HX_tH}}{\|V-V^*\|^2_{X_t^{-1}}\|V-V_*\|^4_H} \leq \liminf_{k\to\infty} \frac{\|V-V_*\|^2_{X_{t_k}^{-1}}}{\log(t_k)} \frac{\|V-V_*\|^2_{HX_{t_k}H}}{\|V-V_*\|^2_{X_{t_k}^{-1}}\|V-V_*\|^4_H}$$
$$= \frac{\liminf_{k\to\infty} \|V-V_*\|^2_{X_{t_k}^{-1}}\|V-V_*\|^2_{HX_{t_k}H}}{\tilde{c}\|V-V_*\|^4_H}\,.$$

Now if we let $H_t = X_t^{-1}/\|X_t^{-1}\|$ and H be the limit of a subset of points $\{H_{t_k}\}$ such that they converge to H, and assume that $\|V-V_*\|_H > 0$, then,

$$\frac{\liminf_{k\to\infty} \|V-V_*\|^2_{X_{t_k}}\|V-V_*\|^2_{HX_{t_k}H}}{\tilde{c}\|V-V_*\|^4_H} = \frac{\liminf_{k\to\infty} \|V-V_*\|^2_{H_{t_k}}\|V-V_*\|^2_{HH_{t_k}^{-1}H}}{\tilde{c}\|V-V_*\|^4_H}$$
$$\leq \frac{\liminf_{k\to\infty} \|V-V_*\|^2_H\|V-V_*\|^2_H}{\tilde{c}\|V-V_*\|^4_H} = \frac{1}{\tilde{c}}\,.$$

Hence,

$$1-\nu \leq \liminf_{t\to\infty} \frac{1}{2\log(t)} \frac{(\Delta_V+\epsilon)\|V-V_*\|^2_{HX_tH}}{\|V-V_*\|^4_H} \leq \frac{(\Delta_V+\epsilon)^2}{2\tilde{c}}$$
$$\implies \limsup_{t\to\infty} \log(t)\|V-V_*\|_{X_t^{-1}} \leq \frac{\Delta_V^2+\epsilon}{2(1-\nu)}\,.$$

Since this holds for any ϵ, we have proved the first part. Note that we have assumed that $\|V-V_*\|_H > 0$. In order to show that this holds in our case, assume otherwise and let $\|V-V_*\|_H = 0$. Then, $H(V-V_*)$ is a vector of 0. But since the kernel of H is the same as H^{-1}, we also have that $H^{-1}(V-V_*) = 0$. Now consider a Λ shifted H. That is let $H_\Lambda = H + \Lambda I_d$. Then, by assumption $H_\Lambda(V-V_*) = \Lambda(V-V_*)$. Furthermore, since we are only considering suboptimal arms, $V-V_*$ is non zero by construction. Hence, $\|V-V_*\|_{H_\Lambda} > 0$. Thus,

$$\frac{\liminf_{k\to\infty} \|V-V_*\|^2_{H_{t_k}}\|V-V_*\|^2_{H_\Lambda H_{t_k}^{-1}H_\Lambda}}{\tilde{c}\|V-V_*\|^4_{H_\Lambda}} = \frac{\liminf_{k\to\infty} \|V-V_*\|^2_{H_{t_k}}\|V-V_*\|^2_{H_{t_k}^{-1}}}{\tilde{c}\|V-V_*\|^4} = 0\,,$$

which is a contradiction because it has to be $0 < 1-\nu < 1$. This finishes part 1 of the proof. In the next part, we connect this result to the spectral norm of a product with features V.

Step 2: Consider any suboptimal arm V. Then, we have to show that

$$\limsup_{t\to\infty} \log(t)\|V\|^2_{X_t^{-1}} \leq \frac{\Delta_V^2}{2(1-\nu)}\,,$$

By part 1, we have that

$$\begin{aligned}
\frac{\Delta_V^2}{2(1-\nu)} &\geq \limsup_{t\to\infty} \log(t)\|V-V_*\|^2_{X_t^{-1}} \\
&= \limsup_{t\to\infty} \log(t)\left((V-V_*)^\top X_t^{-1}(V-V_*)\right) \\
&= \limsup_{t\to\infty} \log(t)\left(V^\top X_t^{-1}V + (V^*)^\top X_t^{-1}(V^*) - 2(V^*)^\top X_t^{-1}V\right) \\
&= \limsup_{t\to\infty} \log(t)\|V\|^2_{X_t^{-1}} + \limsup_{t\to\infty} \log(t)\left((V^*)^\top X_t^{-1}(V^*) - 2(V^*)^\top X_t^{-1}V\right)\,.
\end{aligned}$$

Now first consider $(V^*)^\top X_t^{-1}(V^*)$. By definition, we have that $X_t \succ \mathbb{E}[T_*(t)]V^*(V^*)^\top$. This implies that $X_t^{-1} \prec \frac{1}{\mathbb{E}[T_*(t)]}V^*(V^*)^\top$. Now since, π is a consistent policy with order $\nu > 0$, we have that $\frac{\log(t)}{\mathbb{E}[T_*(t)]} \to 0$ as $T \to \infty$. Otherwise, regret will be linear in T. Hence, $\limsup_{t\to\infty}(V^*)^\top X_t^{-1}(V^*)$ goes to 0.

Next, consider $(V^*)^\top X_t^{-1} V$ and for simplicity assume that V and V^* are perpendicular. Note that the proof is similar for the non perpendicular V as well. Given that V and V^* are perpendicular, we have that $V^\top V^* = 0$. Now let $y = X_t^{-1} V$. Then, clearly, $V = X_t y$. Furthermore, $X_t = \mathbb{E}[T_i(t)] V_i V_i^\top$. Therefore, $V = \mathbb{E}[T_*(t)] V^* (V^*)^\top y + \sum_{i=1,..n, i\neq i^*} \mathbb{E}[T_i(t)] V_i V_i^\top y$. By the perpendicularity assumption between V and V_*, we have that

$$\mathbb{E}[T_*(t)] \|V^*\|^2 (V^*)^\top y + \sum_{i=1,..n, i\neq i^*} \mathbb{E}[T_i(t)] (V^*)^\top V_i V_i^\top y = 0$$
$$\implies (V^*)^\top y = \frac{\sum_{i=1,..n,i\neq i^*} \mathbb{E}[T_i(t)] (V^*)^\top V_i V_i^\top y}{\mathbb{E}[T_*(t)] \|V^*\|^2}.$$

Now,

$$\begin{aligned}\limsup_{t\to\infty} \log(t) (V^*)^\top X_t^{-1} V &= \limsup_{t\to\infty} \log(t) (V^*)^\top y \\ &= \limsup_{t\to\infty} \log(t) \frac{\sum_{i=1,..n,i\neq i^*} \mathbb{E}[T_i(t)] (V^*)^\top V_i V_i^\top y}{\mathbb{E}[T_*(t)] \|V^*\|^2} \\ &= 0\,.\end{aligned}$$

Where the last inequality again follows from the consistency of the policy under consideration. Hence,

$$\limsup_{t\to\infty} \log(t) \|V\|^2_{X_t^{-1}} + \limsup_{t\to\infty} \left((V^*)^\top X_t^{-1} (V^*) - 2 (V^*)^\top X_t^{-1} V \right) = \limsup_{t\to\infty} \log(t) \|V\|^2_{X_t^{-1}} \leq \frac{\Delta_V^2}{2(1-\nu)}$$

which proves the final result. □

Proof of Theorem 2: We will prove the above statement by showing that whenever $|S(u_0, \rho)| < d$, any consistent policy, π, recommends products outside of the customer's feasibility set infinitely often. Note that for any realization of u_0, one can reduce ρ and make it smaller and smaller so that $|S(u_0, \rho)| < d$. Customer disengagement thus follows directly since there is a positive probability, p, of customer leaving the platform whenever a product outside the customer's feasibility set is offered.

In order to show that a consistent policy shows products outside the feasibility set infinitely often, we will use Lemma 1. More specifically, we will construct a counter example such that no consistent policy will satisfy the condition of the Lemma unless it exits the set of feasible products infinitely many times.

Let us assume by contradiction that there exists a policy π that is consistent and offers products inside the feasible set infinitely often. This implies that there exists $\bar{t}$ such that $\forall t > \bar{t}$, $a_t \in \mathcal{S}(u_0, \rho)$. Now under the stated assumptions of Example 1, there are d products in total ($n = d$) and the feature vector of the i^{th} product is the i^{th} basis vector. That is,

$$V_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, V_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, V_3, ..., V_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Further let u_o, the unknown consumer feature vector, and ρ, the tolerance threshold parameter be such that WLOG, $\mathcal{S}(u_0, \rho) = \{2, 3...d\}$ (follows by Definition (2)). That is, only the first product is outside of the feasible set. Also let,

$$R_t^\pi = \begin{bmatrix} T_1^\pi(t) & 0 & \dots \\ \vdots & \ddots & \\ 0 & & T_d^\pi(t) \end{bmatrix},$$

where

$$T_j(t) = \mathbb{E}\left[\sum_{f=1}^{t} \mathbb{1}\{a_f^\pi = j\}\right].$$

$T_j(n)$ is the total number of times the j^{th} product is offered until time t under policy π.
Next consider the following:

$$\begin{aligned}
\limsup_{t\to\infty} \log(t)\|e_1\|^2_{X_t^{-1}} &= \limsup_{t\to\infty} \log(t) e_1^\top X_t^{-1} e_1\,, \\
&= \limsup_{t\to\infty} \log(t) e_1^\top \mathbb{E}\left[\sum_{f=1}^{t} a_f a_f^\top\right]^{-1} e_1 \\
&= \limsup_{t\to\infty} \log(t) e_1^\top \left[R_t\right]^{-1} e_1 \\
&\geq \limsup_{t\to\infty} \log(t) \left(\frac{1}{T_1(t)}\right) \\
&\geq \limsup_{t\to\infty} \log(t) \left(\frac{1}{T_1(\bar{t})}\right) \\
&= \infty\,.
\end{aligned}$$

Where the second to last inequality follows by the fact that $\forall t > \bar{t}$, π recommends products inside the feasible set, $\mathcal{S}(u_0, \rho)$, which does not contain product 1. Furthermore,

$$T_1(\bar{t}) = T_1(\bar{t}+1) = T_1(\bar{t}+2) = ... = \lim_{n\to\infty} T_1(\bar{t}+n).$$

For any finite Δ_{V_1}, and $0 < \nu < 1$, we have that,

$$\limsup_{t\to\infty} \log(t)\|e_1\|^2_{X_t^{-1}} \geq \frac{\Delta_1^2}{2(1-\nu)}.$$

which implies that $\exists a_i$ in the action space such that the condition of Lemma 1 is not satisfied. Hence we have show that there exists no consistent policy that recommends products inside of the feasible set of products infinitely often. Now since ρ is small and p is positive, this leads to a linear rate of regret for all customers. That is,

$$\inf_{\pi\in\Pi^C} \sup_{\rho>0} \mathcal{R}^\pi(T, \rho, p, u_0) = C_1 \cdot T = \mathcal{O}(T)\,,$$

□

Proof of Theorem 3: We prove the result in two parts. In the first part we consider latent attribute realizations for which the optimal aprior product, which is chosen by the GBU policy in the initial round, is not optimal. In this case, if we take the tolerance threshold parameter to be small, there is a positive probability that the customer leaves at the beginning of the time period, which leads to linear regret over this set of customers. In the second part, we consider those customers for which the apriori product is indeed optimal. For these customers, we again take the case when ρ is sufficiently small and reduce the leaving time to the probability of shifting from the first arm to another arm. Since the switched arm is suboptimal and outside of the user threshold, the customer leaves with a positive probability resulting in linear regret for this set of customers.

Recall, by assumption, there are d total products and attribute of the i^{th} product is the i^{th} basis vector. Furthermore, the prior is uninformative. That is, the first recommended product is selected at random. Lets assume WLOG that the GBU policy picks product 1 to recommend. We have two cases to analyze:

- Product 1 is sub optimal for the realized latent attribute vector, u_0.
- Product 1 is optimal for the realized latent attribute vector, u_0.

Lets consider case (i) when product 1 is suboptimal. In this case, if we let ρ to be smaller than the difference between the utility of the optimal product and product 1 ($\rho < u_0^\top (V_* - V_1)$), then the customer leaves with probability p in the current round. Hence, for all such customers

$$\begin{aligned}\mathcal{R}^\pi(T,\rho,p,u_0) &= \sum_{t=1}^{T} r_t(\rho,p,u_0) \\ &\geq \sum_{t=1}^{T} r_t(\rho,p,u_0)\cdot \mathbb{1}(d_1=1)\cdot \mathbb{P}(d_1=1) \\ &\geq T\cdot p \\ &= pT\,.\end{aligned}$$

Thus, for all such customers, the GBU policy incurs at least linear rate of regret.

Next, we consider the customers for which product 1 is optimal. In this case, the customer leaves with probability p when the greedy policy switches from the initial recommendation to some other product. Again, at any such time, t, if we let ρ to be small such that the chosen product is outside of the customer threshold, then we will have disengagement with a constant probability p in that round. This would again lead to linear rate of regret.

Let $E_i^t = \{V_1^\top \hat{u}_t - V_i^\top \hat{u}_t > 0\}$. E_i^t denotes the event that the initially picked product is indeed better than the i^{th} product in the product assortment at time t. Similarly, define G^t to be the event that the GBU policy switches to some other product from product 1 by time t. Then,

$$\begin{aligned}\mathbb{P}(G^t) &= \mathbb{P}\left(\cup_{i=1..n, i\neq i} \cup_{j=1..t} (E_i^j)^c\right) \\ &\geq \mathbb{P}\left((E_i^j)^c\right),\ \forall i = 2,..,n, \forall j = 1,..,t\,.\end{aligned} \tag{4}$$

We will lower bound the probability of product 1 not being the optimal product for any time t under the GBU policy. Since we are dynamically updating the estimated latent customer feature vector, the probability of switching depends on the realization of ε_t, the idiosyncratic noise term that governs the customer response. We will first consider the case of two products ($d = 2$). Furthermore, we will analyse the probability of switching from product 1 to product 2 after round 1 ($(E_2^1)^c$). First note that,

$$\begin{aligned}E_i^t &= \{V_1^t \hat{u}_t - V_i^\top \hat{u}_t \geq 0\} \\ \implies (E_i^t)^c &= \{V_i^t \hat{u}_t - V_1^\top \hat{u}_t > 0\} \\ &= \{V_i^\top \hat{u}_t - V_1^\top \hat{u}_t - V_1^\top u_0 + V_1^\top u_0 - V_i^\top u_0 + V_i^\top u_0 > 0\} \\ &= \{(V_i - V_1)^\top (\hat{u}_t - u_0) > \Delta_i\}\,.\end{aligned}$$

where $\Delta_i = V_1^\top u_0 - V_i^\top u_0$. Now, note that

$$\begin{aligned}\hat{u}_t &= \left[\sum_{f=1}^{t} a_f a_f^\top + \frac{\xi^2}{\sigma^2} I_d\right]^{-1} [a_{1:t}]^\top Y_{f=1:t} \\ \implies \hat{u}_1 &= \begin{bmatrix} 1+\frac{\xi^2}{\sigma^2} & 0 \\ 0 & \frac{\xi^2}{\sigma^2}\end{bmatrix}^{-1} \begin{bmatrix} Y_1 & 0 \\ 0 & 0\end{bmatrix} \\ \implies &= \left[\frac{\sigma^2 Y_1}{\sigma^2+\xi^2}, 0\right]\end{aligned}$$

Therefore, we are interested in the event

$$\left\{\frac{\sigma^2 Y_1}{\sigma^2+\xi^2}<0\right\}=\{Y_1<0\}=\{u_{0_1}+\varepsilon_1<0\}=\{u_{0_1}+\varepsilon_1<0\}=\{\varepsilon_1<-u_{0_1}\}$$

Now note that for any realization of u_0, there is a positive probability of the event above happening. Hence, let

$$\mathbb{P}(\varepsilon_1<-u_{0_1})=C_4>0\,.$$

This implies that $\mathbb{P}(G^t)\geq C_4$. Hence, following the same regret argument as before, we have that for all such customers,

$$\mathcal{R}^{\mathrm{GBU}}(T,\rho,p,u_0)=C_4\cdot T\,.$$

The argument for $d>2$ follows similarly and we skip the details for the sake of brevity. Hence, we have shown that regardless of the realization of the latent user attribute, u_0 the GBU policy incurs linear regret on the customers. That is, $\forall u_0$,

$$\sup_{\rho>0}\ \mathcal{R}^{\mathrm{GBU}}(T,\rho,p,u_0)=C_2\cdot T=\mathcal{O}(T)\,,$$

□

Proof of Theorem 4: We will use the same strategy as in the proof of Theorem 3 with two main exceptions; (i) Because this is the case of no disengagement, we cannot select ρ to be appropriately small. (ii) Since the result is on the expectation of regret over all possible latent attribute realizations, we need to show the result only for a set of customer attributes with positive measure.

Noting (ii) above, we focus on customers for which the first recommended product is suboptimal and show that with positive probability the greedy policy gets "stuck" on this product and keeps on recommending this product. This leads to a linear rate of regret for these customers.

Step 1 (Lower bound on selecting an initial suboptimal product): WLOG assume that product 1 was recommended and consider the set of customers for which u_0 is suboptimal. Note that since u_0 is Multivariate Normal, there is a positive measure of such customers.

Step 2 (Upper bound on the probability of switching from the current product to a different product during the later periods:) Now that we have selected a suboptimal product, we will bound the probability that the GBU policy continues to offer the same product until the end of the horizon. This will lead to a linear rate of regret over all the customers for which the selected product was not optimal.

We will use the same notation as before. Recall that $E_i^t=\{V_1^\top\hat{u}_t-V_i^\top\hat{u}_t>0\}$. E_i^t denotes the event that the initially picked product is indeed better than the i^{th} product in the product assortment at time t. Similarly, G^t denotes the event that the GBU policy switches to some other product from product 1 by time t. Then, we are interested in lower bounding the event that the GBU policy never switches from the product 1 and gets stuck. That is:

$$\begin{aligned}\mathbb{P}((G^t)^c)&=1-\mathbb{P}((G^t))\\&=1-\mathbb{P}\left(\cup_{i=1..n,i\neq i^*}\cup_{j=1..t}(E_i^j)^c\right)\\&\geq 1-\sum_{j=1..t}\sum_{i=1..n,i\neq i^*}\mathbb{P}((E_i^j)^c)\,.\end{aligned}\tag{5}$$

As before, first we consider the case when there are only 2 products. In this case, if we start by recommending product 1, we want to calculate the probability of continuing with Product 1 through out the time horizon. First note that using the same calculation, one can show that if until time t, we continue with only recommending product 1, then the latent attribute estimate at time t is given by

$$\hat{u}_t = \left[\frac{\sigma^2 \sum_{f=1}^{t} Y_f}{t\sigma^2 + \xi^2}, 0 \right].$$

For any time t, we claim that the GBU policy continues to recommend the same product as before if the utility realization at time t is positive. That is, if $Y_{t-1} > 0$ and the GBU policy offered product 1 in rounds $1, ..t-1$ then it will continue recommending product 1 in round t. We prove this claim using induction. Note that the base case of $t = 2$ was proved in the previous proof (reversing the argument in the second part of Theorem 3 results in the base case) and we omit the details here. Now by induction hypothesis, we have that the GBU policy offered product 1 at time $t-1$ because $Y_1, ...Y_{t-2}$ were all positive. Now consider time t let $Y_{t-1} > 0$, Then we have that

$$\hat{u}_{t-1} = \left[\frac{\sigma^2 \sum_{f=1}^{t-1} Y_f}{t\sigma^2 + \xi^2}, 0 \right].$$

We will select product 1 if,

$$\begin{aligned} &\frac{\sigma^2 \sum_{f=1}^{t-1} Y_f}{t\sigma^2 + \xi^2} > 0 \\ &\implies \frac{\sigma^2 \sum_{f=1}^{t-2} Y_f}{t\sigma^2 + \xi^2} + \frac{\sigma^2 Y_{t-1}}{t\sigma^2 + \xi^2} > 0 \end{aligned}$$

But note that by induction hypothesis, the first term of the sum above is positive. Hence, GBU selects product 1 at least when $\frac{\sigma^2 Y_{t-1}}{t\sigma^2+\xi^2} > 0$ which proves the claim. Now note that for any time t, the probability Y_i being positive is independent across time periods. Furthermore,

$$\mathbb{P}\left(\frac{\sigma^2 Y_{t-1}}{t\sigma^2 + \xi^2} > 0 \right) = \mathbb{P}\left(Y_{t-1} > 0\right) = \mathbb{P}\left(u_{0_1} + \varepsilon_t > 0\right)$$

For any t, probability of not switching from the first product is at least

$$\begin{aligned} \mathbb{P}((G^t)^c) = 1 - \mathbb{P}((G^t)^c) &\geq 1 - \sum_{j=1..t} \sum_{i=1..n, i \neq i^*} \mathbb{P}((E_i^j)^c) \\ &= 1 - \sum_{j=1..t} \mathbb{P}\left(\varepsilon_t > -u_{0_1}\right) \\ &= 1 - t\mathbb{P}\left(\varepsilon > -u_{0_1}\right) \end{aligned} \tag{6}$$

Now for any t, if we consider all realIzations of u_0 such that $\mathbb{P}\left(\varepsilon > -u_{0_1}\right) < \frac{1}{t}$, then we have that the above probability is always positive. Note that Product 1 was not optimal, hence, over these customers, the GBU policy incurs linear regret which results in an expected linear regret. That is,

$$\mathbb{E}_{u_0 \sim \mathcal{P}}\left[\mathcal{R}^{GBU}(T, \rho, 0, u_0)\right] = C_3 \cdot T = \mathcal{O}(T).$$

The proof for the case when $d > 2$ follows similarly and we skip the details here for the sake of brevity. Note that unlike Theorem 3, this argument was regardless of disengagement and used the fact that with positive probability, the policy would get stuck on the same arm with which it started regardless of what the real optimal is. $\square$

C. Upper Bound for Constrained Bandit

Proof of Theorem 5: Consider any feasible $\rho > 0$ and let $\tilde{\gamma}$ be such that only a single product remains in the constrained exploration set. Note that a feasible γ that ensures that only a single product is chosen for exploration is $\gamma < \frac{1}{\sqrt{2}}$. Such a selection would ensure that $\mathbf{OP}(\gamma)$ picks a single product ($\tilde{i}$) in the exploration phase. Now let $\tilde{\gamma} = \frac{1}{\sqrt{2}}$ and consider

$$\mathcal{W}_{\lambda,\tilde{\gamma}} := \{u_0 : V_{\tilde{i}}^\top u_0 > \max_{i-1,..,n, i\neq\tilde{i}} V_i^\top u_0\}.$$

Then we have that $\forall u_0 \in \mathcal{W}_{\lambda,\tilde{\gamma}}$, customers are going to continue engaging with the platform since the recommended product is the corresponding optimal product. Next, since the prior is a multivariate normal, we have that $\mathbb{P}(\mathcal{W}_{\lambda,\tilde{\gamma}}) > 0$. This holds because by assumption since V_i is the i^{th} basis vector and u_0 is multivariate normal with prior mean of 0 across all dimensions. So, the probability of sampling a u_0 such that $u_{0_{\tilde{i}}} > u_{0_j},\ \forall j = 1,..,d,\ j \neq \tilde{i}$ has a positive measure under the prior assumption. We claim that for any ρ, the regret incurred from this policy will be optimal. Consider two cases: (i) When ρ is such that their is more than 1 product within the customer's relevance threshold. That is, $|\mathcal{S}(u_0, \rho)| > 1$ (ii) When there is a single product within the customer's tolerance threshold, ρ. That is, $|\mathcal{S}(u_0, \rho)| = 1$. In both cases, $\tilde{i}$, which is the only product in the exploration phase, is contained in $|\mathcal{S}(u_0, \rho)|$. That is, $\forall u_0 \in \mathcal{W}_{\lambda,\tilde{\gamma}},\ \tilde{i} \in \mathcal{S}(u_0, \rho)$. Hence, there are no chances of customer disengagement if product $\tilde{i}$ is offered to the customer. Furthermore, regret over all such customers is infact 0 since the platform recommends their optimal product. This proves the result. □

Proof of Theorem 6: We will prove the above result in three steps. In the first step we will lower bound the probability that the constrained exploration set, Ξ, contains the optimal product for an incoming vector. In the second step we will lower bound the probability of customer engagement over the constrained set. Finally, in the last step, we use the above lower bounds on probabilities to upper bound regret from the Constrained Bandit algorithm.

Step 1 (Lower bounding the probability of not choosing the optimal product for an incoming customer in the constrained set): Let, $\mathcal{E}_{no-optimal}$, be the event that the optimal product, V_* for the incoming user is not contained in Ξ. Also let $\tilde{i} = \arg\max_{V\in[-1,1]^d}\ \bar{u}^\top V$, denote the attributes of the prior optimal product. Notice that $V_{\tilde{i}} = \bar{u}$ since $\|\bar{u}\|_2 = 1$. Also recall that

$$V_* = \arg\max_{V\in[-1,1]^d}\ u_0^\top V,$$

denotes the current optimal product which is unknown because of unknown customer latent attributes. We are interested in

$$\mathbb{P}(\mathcal{E}_{no-optimal}) = \mathbb{P}(V_* \notin \Xi).$$

In order to characterize the above probability, we focus on the structure of the constrained set, Ξ. Recall that Ξ is the outcome of Step 1 of Constrained Bandit (Algorithm 2) and uses $\mathbf{OP}(\gamma)$ to restrict the exploration space. It is easy to observe that Ξ in the continuous feature space case would be centred around the prior optimal product vector ($\bar{u}$) and will contain all products that are at most γ *away* from each other. Figure 6 plots the constrained set under consideration. We are interested in characterizing the probability

of the event that $u_0 \notin [\bar{u}_l, \bar{u}_r]$ where $\bar{u}_l$ and $\bar{u}_r$ denote the attributes of the farthest products inside a γ constrained sphere. Furthermore, $\bar{u}_l$ and $\bar{u}_r$ are divided in two equal halves of length $\frac{\gamma}{2}$ by a perpendicular line segment between $\bar{u}$ and the origin due to the symmetry of the constrained set. Simple geometric analysis yields that $\bar{u}$ and $\bar{u}_l$ are $\sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)}$ apart. The distance between $\bar{u}$ and $\bar{u}_r$ follows symmetrically.

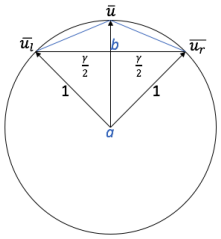

Figure 6 Analyzing the probability of $\mathcal{E}_{no-optimal}$. Note that we are interested in characterizing the length of line segments $(\bar{u}_l, \bar{u})$ and $(\bar{u}, \bar{u}_r)$

Having calculated the distance between $\bar{u}$ and $\bar{u}_l$, we are now in a position to characterize the probability of $\mathcal{E}_{no-optimal}$.

$$\mathbb{P}\left(\mathcal{E}_{no-optimal}\right) = \mathbb{P}\left(V_* \notin \Xi\right) = \mathbb{P}\left(\|u_0 - \bar{u}\|_2 \geq \sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)}\right).$$

Note by Holder's inequality that,

$$\sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)} \leq \|u_0 - \bar{u}\|_2 \leq \|u_0 - \bar{u}\|_1,$$

which implies that,

$$\mathbb{P}\left(\mathcal{E}_{no-optimal}\right) = \mathbb{P}\left(\|u_0 - \bar{u}\|_2 \geq \sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)}\right) \leq \mathbb{P}\left(\|u_0 - \bar{u}\|_1 \geq \sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)}\right).$$

Note that $u_0 \sim \mathcal{N}(\bar{u}, \frac{\sigma^2}{d^2} I_d)$. Hence, using Lemma 2 in Appendix E, we have that,

$$\mathbb{P}\left(\|u_0 - \bar{u}\|_1 \leq \sqrt{2\left(1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}\right)}\right) \geq 1 - 2d\exp\left(-\left(\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\right).$$

Hence,

$$\mathbb{P}\left(\mathcal{E}_{no-optimal}\right) \leq 2d\exp\left(-\left(\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\right).$$

Step 2 (Lower bounding the probability of customer disengagement due to relevance of the recommendation): Recall that customer disengagement decision is driven by the relevance of the recommendation and the tolerance threshold of the customer. Hence,

$$\mathbb{P}(u_0^\top V_* - u_0^\top V_i < \rho) = \mathbb{P}(u_0^\top u_0 - u_0^\top u_i < \rho | u_0, u_i \in \Xi)$$

$$
\begin{aligned}
&= \mathbb{P}(u_0^\top (u_0 - u_i) < \rho | u_0, u_i \in \Xi) \\
&\geq \mathbb{P}\left(\|u_0\|_2 < \frac{\rho}{\gamma} | u_0, u_i \in \Xi \right) \\
&\geq \left(1 - 2d \exp\left(- \left(\frac{\frac{\rho}{\gamma} - \sum_{i=1}^{i=d} \bar{u}_i}{\sigma} \right)^2 \right) \right).
\end{aligned}
$$

where the last inequality follows by Lemma 2. This in-turn shows that with probability at least $\left(1 - 2d \exp\left(- \left(\frac{\frac{\rho}{\gamma} - \sum_{i=1}^{i=d} \bar{u}_i}{\sigma} \right)^2 \right) \right)$, customers will not leave the platform because of irrelevant product recommendations. We let such latent attribute realizations be denoted by the event $\mathcal{E}_{relevant}$.

Step 3 (Sub-linearity of Regret): Recall that

$$
r_t(\rho, p, u_0) = \begin{cases} u_0^\top V_* & \text{if } \ d_{t'} = 1 \ \text{ for any } \ t' < t\,, \\ u_0^\top V_* - u_0^\top a_t & \text{otherwise.} \end{cases}
$$

In other words,

$$
r_t(\rho, p, u_0) = (u_0^\top V_* - u_0^\top a_t) \mathbb{1}\{L_{t,\rho,p} = 1\} + u_0^\top V_* \mathbb{1}\{L_{t,\rho,p} = 0\}.
$$

Nevertheless,

$$
\begin{aligned}
r_t(\rho, p, u_0) &= \left(u_0^\top V_* - u_0^\top a_t\right) \mathbb{1}\{L_{t,\rho,p} = 1\} + (u_0^\top V_*) \mathbb{1}\{L_{t,\rho,p} = 0\} \\
&= \left(u_0^\top V_* - u_0^\top a_t\right) \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\} + (u_0^\top V_*)(1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\}) \\
&= \left(u_0^\top V_* - u_0^\top a_t\right) \Pi_{t=1}^\top \mathbb{1}\{dt = 0\} + (u_0^\top V_* + u_0^\top a_t - u_0^\top a_t)(1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\}) \\
&= (u_0^\top V_* - u_0^\top a_t) + u_0^\top a_t (1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\}).
\end{aligned}
$$

Note that the first part in the above expression is related to the regret of the classical bandit setting where the customer does not disengage while the second part is associated with the regret when the customer disengages from the platform and the platform incurs maximum regret.

Next, focusing on cumulative regret and taking expectation over the random customer response on quality feedback (ratings), we have that,

$$
\begin{aligned}
\mathbb{E}_{U_0 \sim \mathcal{P}}\left[\mathcal{R}^{CB}(T, \rho, p, u_0)\right] = \mathbb{E}_{U_0 \sim \mathcal{P}}\left[\sum_{t=1}^{T} r_t(\rho, p, u_0)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{\top} (u_0^\top V_* - u_0^\top a_t) + u_0^\top a_t (1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\})\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[(u_0^\top V_* - u_0^\top a_t)\right] + \mathbb{E}\left[u_0^\top a_t \left(1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\}\right)\right].
\end{aligned}
$$

Note that conditional on fraction w of customers, we have that these customers would never disengage from the platform due to irrelevant personalized recommendations. Hence,

$$
1 - \Pi_{t=1}^\top \mathbb{1}\{d_t = 0\} = 0\,,
$$

since there is no probability of leaving when a product within the constraint set is recommended and meets the tolerance threshold (w, analyzed in the previous step). Hence,

$$
\mathcal{R}^{CB(\lambda,\gamma)}(T, \rho, p, u_0 | u_0 \in \mathcal{E}_{relevant}) = \sum_{t=1}^{T} (u_0^\top V_* - u_0^\top a_t).
$$

Now notice that for any realization of u_0, Lemma 3 of Appendix E shows that

$$\mathcal{R}^{CB(\lambda,\gamma)}(T,\rho,p,u_0|u_0\in\mathcal{E}_{relevant}) \leq 4\sqrt{Td\,\log\left(\lambda+\frac{TL}{d}\right)}\left(\sqrt{\lambda}S+\xi\sqrt{2\log\frac{1}{\delta}+d\log\,(1+\frac{TL}{\lambda d})}\right).$$

with probability at least 1-δ if $\|u_0\|_2 \leq S$. From Step 2, we have that all w fraction of customers have $\|u_0\|_2 \leq \frac{\rho}{\gamma}$. Hence first we replace S with $\frac{\rho}{\gamma}$. Finally, letting $\delta = \frac{1}{\sqrt{T}}$, we get that

$$\begin{aligned}\mathcal{R}^{CB(\lambda,\gamma)}(T,\rho,p,u_0|u_0\in\mathcal{E}_{relevant}) &\leq 4\sqrt{Td\log\left(\lambda+\frac{TL}{d}\right)}\left(\sqrt{\lambda}\frac{\rho}{\gamma}+\xi\sqrt{\log(T)+d\log\left(1+\frac{TL}{\lambda d}\right)}\right)+\frac{1}{\sqrt{T}}T\\ &= \tilde{\mathcal{O}}\left(\sqrt{T}\right).\end{aligned}$$

Recall that ξ is the error sub-Gaussian parameter and λ is the L2 regularization parameter. Rearranging the terms above gives the final answer. □

D. Selecting set diameter γ

In the previous section, we proved that the Constrained Bandit algorithm achieves sublinear regret for a large fraction of customers. This fraction depends on the constrained threshold tuning parameter γ and other problem parameters (see Theorem 6). In this section, we explore this dependence in more detail and provide intuition on the selection of γ that maximizes this .

Recall, from Theorem 6, that the fraction of customers who remain engaged with the platform is lower bounded by,

$$w \;=\; \left(1-2d\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\right)\left(1-2d\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)\right).$$

This fraction comprises of two parts. The first part, $\left(1-2d\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\right)$, denotes the fraction of customers for which the corresponding optimal product is contained in the constrained exploration set, Ξ. Notice that the fraction of customers for which the optimal product is contained in the constrained set increases as the constraint threshold, γ, increases. This follows since a larger γ implies a larger exploration set and more customer that can be served with their most relevant recommendation. Similarly, the second part, $\left(1-2d\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)\right)$, denotes the fraction of customers who will not disengage from the platform due to irrelevant recommendations in the learning phase. Contrary to the previous case, as the constraint threshold γ increases, the fraction of customers guaranteed to engage decreases. Intuitively, as the exploration set becomes larger, there is wider range of offerings with more variability in the relevance of the recommendations for a particular customer. This wider relevance in turn leads to a decrease in the probability of engagement of a customer. Hence, γ can either increase or decrease the fraction of engaged customers based on the other problem parameters.

In Figure 7, we plot the fraction of customers who will remain engaged with the platform as a function of the set diameter, γ, for different values of tolerance threshold, ρ. As noted earlier, the fraction of engaged customers is not monotonically increasing in γ. When γ is small, the constrained set for exploration (from

Step 1 of Algorithm 2) is over constrained. Hence, increasing γ leads to an increase in the fraction of engaged customers. Nevertheless, increasing it above a threshold implies that customers are more likely to disengage from the platform due to irrelevant recommendations. Hence, increasing γ further leads to a decrease in the fraction of engaged customers. We also note that as customers become less quality concious (small ρ), the fraction of engaged customers increases for any chosen value of γ. This again follows from the fact that a higher value of ρ implies a higher probability of customer engagement in the learning phase. This increase in engagement probability during the learning phase encourages less conservative exploration (larger γ). The

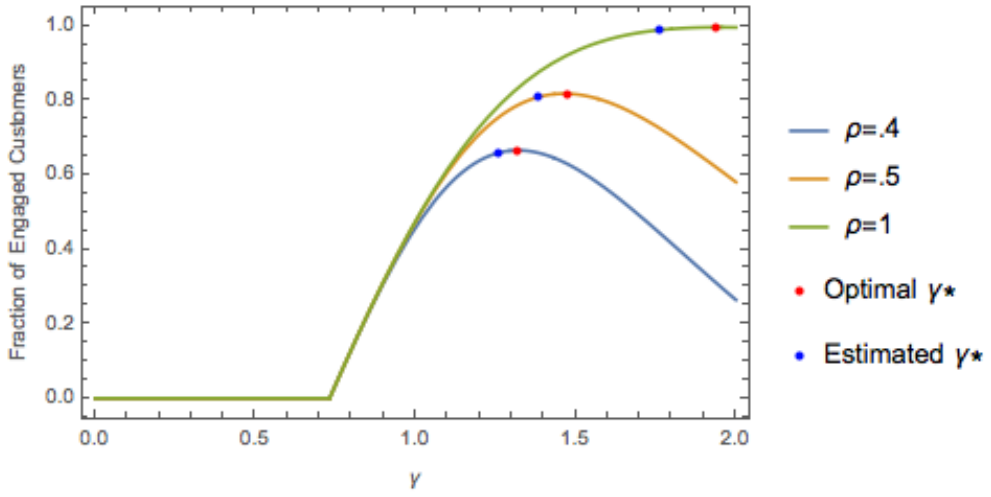

Figure 7 Fraction of engaged customer as a function of the set diameter γ for different values of tolerance threshold, ρ. A higher ρ implies that the customer is less quality concious. Hence, for any γ, this ensures higher chance of engagement. We also plot the optimal γ that ensures maximum engagement and an approximated γ that can be easily approximated. The approximated γ is considerably close to the optimal γ and ensures high level of engagement.

above discussion alludes to the fact that the optimal γ that maximizes the fraction of engaged customers is a function of different problem parameters and is hard to optimize in general. Nevertheless, consider the following:

$$
\begin{aligned}
w \;=\; & \left(1-2d\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\right)\left(1-2d\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)\right) \\
=\; & 1-2d\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)-2d\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)+ \\
& 4d^2\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right)\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)
\end{aligned}
$$

$$\approx \frac{1}{2d^2} - \frac{1}{2d}\exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right) - \frac{1}{2d}\exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right)$$

Hence, in order to maximize w, we have to solve the following minimization problem:

$$\min_{\gamma}\ \exp\left(-\frac{1-\sqrt{\left(1-\frac{\gamma^2}{4}\right)}}{\sigma}\right) + \exp\left(-\left(\frac{\frac{\rho}{\gamma}-\sum_{i=1}^{i=d}\bar{u}_i}{\sigma}\right)^2\right). \tag{7}$$

While Problem (7) has no closed form solution, we consider the following problem:

$$\min_{\gamma}\ \frac{1}{\sigma}\sqrt{\left(1-\frac{\gamma^2}{4}\right)} - \frac{\rho^2}{\gamma^2\sigma^2}. \tag{8}$$

Note that (8) is an approximation of (7) based on the Taylor series expansion of the exponent function and assuming that the joint term in the second exponent will be sufficiently small. Solving (8) using FOC conditions, a suitable choice of γ yields the following:

$$\gamma^* \in \left\{\gamma : \rho = \frac{\sqrt{\sigma}\gamma^2}{2\left(4-\gamma^2\right)^{1/4}} \text{ and } \gamma > 0\right\}.$$

While γ^* is not optimal, it provides directional insights to managers on suitable choices of γ. For example, as ρ increases the estimated optimal γ also increases. Furthermore, it decreases with the prior variance, σ. A lower variance yields better understanding of the unknown customer and leads to lower size of the optimal exploration set. Similarly, as the latent vector dimension, d, increases, there are higher chances of not satisfying customer relevance thresholds in the learning phase. This leads to a more constrained exploration.

In order to analyze the estimated optimal γ, we compare the estimated optimal γ with the numerically calculated optimal γ for different values of ρ, the customer tolerance threshold. In Table 3, we show the gap in the lower bound of engaged customers from choosing the optimal γ vs the estimated γ. Note that the approximated optimal γ performs well in terms of the fraction of engaged customers. More specifically, the estimated γ loses at most 1% customers because of the approximation.

Tolerance Threshold (ρ)	***Optimal γ^****	***Estimated γ^****	***% Gap in Engagement***
0.4	1.31	1.25	1.1%
0.5	1.47	1.37	1.1%
1.0	1.93	1.76	0.07%

Table 3 Optimal vs. estimated γ threshold for different values of customer tolerance threshold, ρ. Note that the % gap between the lower bound on engaged customers is below 1.1% showing that the estimated γ is near optimal.

We note that this optimal selection of γ is based on the model setting of Theorem 6. In Section 6, we discuss the selection of γ for more general settings.

E. Supplementary Results

LEMMA 2. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(\mu, \sigma^2 I)$ be a multivariate normal random variable with mean vector $\mu \in \mathbb{R}^d$. Let $S \in \mathbb{R}^d$ be such that $S \geq \sum_{i=1}^{i=d} \mu_i$. Then,*

$$\mathbb{P}(\|X\|_1 \leq S) \geq 1 - 2d \exp\left(-\left(\frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right)^2\right)$$

Proof: First note that,

$$\|X\|_1 = \sum_{i=1}^{i=d} |X_i| = \sum_{i=1}^{i=d} \sigma\left(\frac{|X_i - \mu_i + \mu_i|}{\sigma}\right) \leq \sum_{i=1}^{i=d} \sigma\left(\frac{|X_i - \mu_i|}{\sigma} + \frac{\mu_i}{\sigma}\right)$$

Then, we have that

$$\begin{aligned}
\mathbb{P}(\|X\|_1 > S) &\leq \mathbb{P}\left(\sum_{i=1}^{i=d} \sigma\left(\frac{|X_i - \mu_i|}{\sigma} + \frac{\mu_i}{\sigma}\right) \geq S\right) \leq \mathbb{P}\left(\sum_{i=1}^{i=d} \frac{|X_i - \mu_i|}{\sigma} \geq \frac{S - \sum_{i=1}^{i=d} \mu_i}{\sigma}\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{i=d} |Z| \geq \frac{S - \sum_{i=1}^{i=d} \mu_i}{\sigma}\right) \leq d\left(\mathbb{P}\left(|Z| \geq \frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right)\right) \\
&= d\left(\mathbb{P}\left(Z \geq \frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right) + \mathbb{P}\left(Z \leq -\frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right)\right) \\
&= 2d\left(\mathbb{P}\left(Z \geq \frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right)\right) \\
&\leq 2d \exp\left(-\left(\frac{S - \sum_{i=1}^{i=d} \mu_i}{d\sigma}\right)^2\right)
\end{aligned}$$

where $Z \in \mathbb{R}^1 \sim \mathcal{N}(0,1)$ and the first set of inequalities follow by the pigeon-hole principle and the union bound. The last inequality follows by the tail probability of standard normal random variables. The result follows easily. □

LEMMA 3 **(Theorem 2 in Abbasi-Yadkori et al. (2011))**. *Let $\mathcal{F}_{t=0}$ be a filtration. Let $\eta_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$ measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e.*

$$\forall \lambda \in \mathbb{R}, E[e^{\lambda \eta} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right)$$

Let $X_{t=1}^{t \to \infty}$ be an $\mathbb{R}^d$ valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$V_t = V + \sum_{s=1}^{s=T} X_s X_s^\top, S_t = \sum_{s=1}^{s=t} \eta_s X_s$$

Then for any $\delta > 0$, the following holds with probability at least 1-δ for all $t \geq 0$.

$$\|S_t\|_{\bar{V}_t^{-1}} \leq 2R^2 \log\left(\frac{det(\bar{V}_t)^{1/2} det(\lambda I)^{-1/2}}{\delta}\right) \tag{9}$$

Now let $V = I\lambda$, $\lambda > 0$, define $Y_t = X_t^\top \theta + \eta_t$ and assume that $\|\theta_*\|_2 \leq S$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$, θ^* lies in the set

$$\mathcal{C}_t : \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\bar{V}_t} \leq \left(R\sqrt{dlog\left(\frac{1 + tL^2}{\delta}\right)} + \sqrt{\lambda} S\right)\right\} \tag{10}$$

LEMMA 4 **(Lemma 5 in Lattimore and Szepesvari (2016))**. *Let $\mathbb{P}$ and $\mathbb{P}'$ be measures on the same measurable space $(\Omega, \mathcal{F})$. Then for any event $A \in \mathcal{F}$,*

$$\mathbb{P}(A) + \mathbb{P}'(A^c) \geq \frac{1}{2} \exp\left(-\mathrm{KL}(\mathbb{P}, \mathbb{P}')\right)$$

where A^c is the complementer event of A ($A^c = \Omega \setminus A$) and $\mathrm{KL}(\mathbb{P}, \mathbb{P}')$ is the relative entropy between $\mathbb{P}$ and $\mathbb{P}'$.

LEMMA 5 **(Lemma 6 in Lattimore and Szepesvari (2016))**. *Let $\mathbb{P}$ and $\mathbb{P}'$ be probability measures on $(a_1, Y_1, ...a_n, Y_n) \in \Omega_n$ for a fixed bandit policy π interacting with a linear bandit with standard Gaussian noise and parameter u_0 and u_0' respectively. Under these conditions, the KL divergence of $\mathbb{P}$ and $\mathbb{P}'$ can be computed exactly and is given by*

$$\mathrm{KL}(\mathbb{P}, \mathbb{P}') = \frac{1}{2} \sum_{j \in \{1,\ldots,n\}} \mathbb{E}\left[T_j(t)\right] \left(V_j^\top (u_0 - u_0')\right)^2,$$

where

$$T_j(T) = \sum_{t=1}^{T} \mathbb{1}\{a_t^\pi = j\}.$$

$T_j(n)$ is the total number of times the j^{th} product is offered until time T under policy π.