Learning Personalized 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

Problem definition: We study personalized product recommendations on platforms when customers have unknown preferences. Importantly, customers may *disengage* when offered poor recommendations.

Academic / Practical Relevance: Online platforms often personalize product recommendations using bandit algorithms, which balance an exploration-exploitation tradeoff. However, customer disengagement—a salient feature of platforms in practice—introduces a novel challenge, since exploration may cause customers to abandon the platform. We propose a novel algorithm that constrains exploration to improve performance.

Methodology: We present evidence of customer disengagement using data from a major airline's ad campaign; this motivates our model of disengagement, where a customer may abandon the platform when offered irrelevant recommendations. We formulate the customer preference learning problem as a generalized linear bandit, with the notable difference that the customer's horizon length is a function of past recommendations.

Results: We prove that no algorithm can keep *all* customers engaged. Unfortunately, classical bandit algorithms provably over-explore, causing *every* customer to eventually disengage. Motivated by the structural properties of the optimal policy in a scalar instance of our 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 by keeping a significant fraction of customers engaged.

Managerial Implications: Platforms should be careful to avoid over-exploration when learning customer preferences if customers have a high propensity for disengagement. Numerical experiments on movie recommendations data demonstrate that our algorithm can significantly improve customer engagement.

Key words: bandits, recommendation systems, collaborative filtering, disengagement, cold start

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 rank (*i.e.,* 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 sequentially offer random products to a customer 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 the customer is 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 limited to *intentional* exploration: we show that a greedy exploitation-only algorithm also underperforms due to excessive natural exploration. 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. *Evidence of disengagement:* Using panel data on ad campaigns from a major airline, we show that the relevance of past recommendations affects a customer's decision to stay on the platform.

2. *Disengagement model:* We modify the classical generalized linear bandit formulation for making personalized product recommendations, so that 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 first show that no algorithm can keep *every* customer engaged; however, we can hope to perform well on a subset of customers. Unfortunately, classical bandit and greedy algorithms over-explore, causing every customer to eventually disengage.

4. *Algorithm:* We first reduce the scalar instance of our problem to a known scheduling problem, implying that it has an optimal index-based policy. We analyze the structural properties of this policy, and show that it avoids arms that are likely sub-optimal for the entire time horizon.

Motivated by this result, we propose the Constrained Bandit algorithm, which modifies standard bandit strategies by constraining the product set upfront using a novel integer programming formulation. The integer program leverages information on other customers on the platform to select a subset of products that are likely to be relevant for the incoming customer. Unlike classical approaches, the Constrained Bandit achieves sublinear regret for a significant fraction of customers.

5. *Numerical experiments:* Extensive numerical experiments on synthetic and real world movie recommendation data demonstrate that the Constrained Bandit significantly improves both regret and the length of time that a customer is engaged with the platform.

1.2. Related Literature

The value of personalizing the customer experience has been recognized for a long time (Surprenant and Solomon 1987), with recent focus on online content and product recommendations (Besbes et al. 2015, Demirezen and Kumar 2016, Farias and Li 2019). We take the widely-used collaborative filtering framework (Sarwar et al. 2001, Su and Khoshgoftaar 2009) as our point of departure. All these methods suffer from the cold start problem (Schein et al. 2002): when a new customer arrives, no data is available on her product preferences, 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. These algorithms essentially 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. 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. Instead, we *transfer* knowledge (Bastani 2021) from other customers on the platform to make more palatable recommendations from the beginning.

At a high level, our work also relates to the broader literature employing bandits on platforms, such as assortment selection (Agrawal et al. 2016, Kallus and Udell 2016) or matching heterogeneous workers to jobs (Johari et al. 2017). 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. These studies rely on optimally balancing the exploration-exploitation tradeoff under bandit feedback. A closely related goal is best arm identification, where one seeks to maximize the probability of identifying the best arm at the end of the time horizon (Garivier and Kaufmann 2016). More generally, in product ranking and selection, one may wish to learn the top k products in order to recommend *assortments*. Suitable algorithms offer customers subsets of products to infer the underlying customer choice model (Chen et al. 2018, Feng et al. 2018).

A key assumption underlying the literature above 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 different analysis.

Customer Disengagement: Customer disengagement and its relation to service relevance have been extensively studied. For instance, Venetis and Ghauri (2004) use a structural model to establish that service relevance contributes to long term customer relationships and retention. Bowden (2009) models the differences in engagement behavior across new and repeat customers. Sousa and Voss (2012) study the impact of e-service relevance 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 find similarly that poor recommendations result in customer disengagement on airline campaign data. 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 relevance 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 relevance. None of these papers study learning in the presence of customer disengagement.

A notable exception is Johari and Schmit (2018), who learn 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 historical product ratings/choices. 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 machine learning techniques (collaborative filtering and bandits) and the extensive modeling literature on customer disengagement and service relevance optimization.

2. Motivation

We use customer panel data from a major commercial airline, obtained as part of a 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.

The airline conducted 7 large-scale *non-targeted* ad campaigns over the course of a year. Each campaign emailed loyalty customers destination recommendations hand-selected by a marketing team at discounted rates; all customers received the same recommendations. 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 has already disengaged with the airline campaign and is no longer responding to recommendations.

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) relevance 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 relevance of the recommendation in the 6^{th} (previous) campaign affected the customer's response in the 7^{th} (current) campaign after controlling for the relevance 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 relevance of the current campaign's recommendations; if we instead find that the relevance 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 estimate a personalized relevance score of recommendations for each customer by applying collaborative filtering to click data from the first 5 campaigns (see Appendix A.1 for details).[1] While we do not have access to other customer-specific variables (e.g., customer loyalty tier/status, unobserved travel preferences), we show in Appendix A.2 that we still obtain unbiased treatment effect estimates because the campaigns were pre-determined and non-targeted.

Regression Specification. We perform our regression over the 7^{th} (current) campaign's click data. Specifically, we examine if the relevance of the recommendation in the 6^{th} (previous) campaign affected the customer's response in the current campaign after controlling for the relevance 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 logistic regression:

$$\mathbb{P}(y_i = 1) = f(\beta_0 + \beta_1 \cdot prev_i + \beta_2 \cdot curr_i)\,,$$

where f is the logistic function and y_i is the click outcome for customer i in the current campaign. We fit an intercept term β_0, the effect of the previous campaign's recommendation relevance on the customer's click likelihood β_1, and the effect of the current campaign's recommendation relevance 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 relevance.

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.,* $\beta_1 > 0$ (p-value $= 7 \times 10^{-9}$). We caution that these results are based on observational data, and are therefore suggestive rather than definitive.

[1] A version of this score was used by the airline in a live pilot for making personalized recommendations to customers in similar ad campaigns.

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

We motivate our formulation by embedding it within the popular *collaborative filtering* framework (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 customer i's utility from product j. Most entries in this matrix are missing since a typical customer has only evaluated a small subset of available products. Collaborative filtering uses a low-rank decomposition, $A = U^\top V$ (in the linear utility case), where $U \in \mathbb{R}^{d\times m}, V \in \mathbb{R}^{d\times n}$ 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 $U_i \in \mathbb{R}^d$ (column 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 utility of product j to customer i is $U_i^\top V_j$. 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 $\{V_j\}_{j=1}^n$. All product attributes are bounded, *i.e.,* there exists $L > 0$ such that $\|V_i\|_2 \le L$ is satisfied for all $i \in \{1, ..., m\}$.

New Customer: 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 model $U_0 \sim \mathcal{P}$, 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 analytical tractability, we take $\mathcal{P}$ to be a multivariate normal distribution $\mathcal{N}(0, \sigma^2 I_d)$.

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 expected utility $U_0^\top a_t$. More generally, we can model nonlinear customer utilities using a generalized linear model, *i.e.,*

$$\mu\left(U_0^\top a_t\right) + \varepsilon_t\,,$$

where ε_t is independent, zero-mean ξ-subgaussian noise and the link function μ is strictly increasing. For instance, in linear regression, we have continuous outcomes with $\mu(x) = x$; in logistic regression, we have binary outcomes with $\mu(x) = \exp(x)/(1+\exp(x))$; in Poisson regression, we have integer-valued outcomes $\mu(x) = \exp(x)$. We seek to learn U_0 through the customer's feedback from a series of recommendations in order to eventually offer her the best available product

$$V_* = \underset{V_j \in \{V_1,\ldots,V_n\}}{\arg\max} \; \mu\left(U_0^\top V_j\right) = \underset{V_j \in \{V_1,\ldots,V_n\}}{\arg\max} \; U_0^\top V_j \,.$$

We impose that $\mu\left(U_0^\top V_*\right) > 0$, *i.e.,* the customer receives positive expected utility from being matched to her most preferred product on the platform.

In the case of a *nonlinear* link function, we make some additional assumptions from the generalized linear bandit literature (Filippi et al. 2010). Specifically, we impose that our link function μ is k_μ-Lipschitz continuous and continuously differentiable with $\mu'(\cdot) \geq c_\mu$ on its domain. Furthermore, the magnitude of the customer's utility is non-negative and bounded by Y_{max} almost surely.

The problem of learning U_0 now reduces to a classical generalized linear bandit, 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 regret definition depart from the standard generalized linear bandit by modeling customer disengagement.

3.1. 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 $\Upsilon_t \in \{0,1\}$, where $\Upsilon_t = 1$ signifies that the customer has disengaged (and receives zero utility for the remainder of the time horizon T) and $\Upsilon_t = 0$ signifies that the customer has chosen to remain engaged for the next time step.

There are many ways to model disengagement. Our primary model is loosely inspired by the experimental results of Fitzsimons and Lehmann (2004), who find that irrelevant recommendations can lead customers to ignore future recommendations due to the activation of a reactance state. For each customer, we model 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:

$$\Pr[\Upsilon_t = 1 \mid a_t] = \begin{cases} 0 & \text{if } \; u_0^\top a_t \; \geq \; \rho \,, \\ p & \text{otherwise.} \end{cases}$$

In other words, each customer is satisfied with an expected utility of at least $\mu^{-1}(\rho)$ from a recommendation. If the platform makes a recommendation that results in a utility less than this threshold, the customer will disengage with some positive probability $p > 0$. Here, $\rho < u_0^\top V_*$, *i.e.,* there is at least one product on the platform that is acceptable to the customer. Note that we recover the classical linear bandit formulation (with no disengagement) when $\rho \to -\infty$. We discuss alternative disengagement models in Section 3.4, and obtain qualitatively similar results.

We seek to construct a non-anticipating sequential policy $\pi = \{a_1, \cdots, a_T\}$ that learns U_0 over time to maximize the customer's utility on the platform. Non-anticipating policies $\Pi : \pi = \{\pi_t\}$ form a sequence of random functions π_t that depend only on observations until time t.

Remark 1. All policies assume knowledge of the tolerance parameter ρ, the disengagement propensity p, and the distribution of latent customer attributes $\mathcal{P}$. In practice, these quantities 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 §5). Furthermore, the disengagement parameters ρ and p may vary by customer; in Appendix C.3, we extend to the case where each customer's disengagement parameters are sampled from a known joint distribution.

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 $\sqrt{V^\top X V}$. 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 1. *Let $\mathcal{S}(u_0, \rho)$ be the subset of products that satisfy the tolerance threshold for a customer with latent attribute vector u_0. More specifically,*

$$\mathcal{S}(u_0, \rho) := \{i : \ u_0^\top V_i \ \geq \ \rho\} . \tag{1}$$

In the classical bandit, 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 increases, $|\mathcal{S}(u_0, \rho)|$ decreases.

3.2. Performance Metric

Typically, the performance of π is measured by its cumulative expected regret (Lai and Robbins 1985). In particular, we would compare the performance of our policy π against an oracle policy π^* that knows U_0 in advance and always offers the customer's preferred product V_*. At time t, the instantaneous expected regret of policy π for a new customer with realized attributes $U_0 = u_0$ is:

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

This is 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 cumulative expected regret for a given customer is then

$$\mathcal{R}^\pi(T, \rho, p, u_0) = \sum_{t=1}^{T} r_t^\pi(\rho, p, u_0)\,. \tag{2}$$

The usual goal is to find a policy π that minimizes the cumulative expected regret for a new customer whose attributes are sampled from $\mathcal{P} = \mathcal{N}(0, \sigma^2 I_d)$. However, it is easy to show that no policy can obtain sublinear regret over *all* customers when disengagement is salient.

Proposition 1 **(Hardness Result)**. *When $\rho < \infty$, any non-anticipating policy $\pi \in \Pi$ cannot achieve sublinear regret for all customers. That is, $\forall T$,*

$$\inf_{\pi \in \Pi} \mathcal{R}^\pi(T, \rho, p, u_0) \;=\; \Omega(T)\,.$$

The proof is given in Appendix B.1. In other words, regardless of the policy chosen, there exists a subset of users (with positive measure under $\mathcal{P}$) who incur linear regret. Proposition 1 shows that product recommendation with customer disengagement *requires* making a trade-off over the types of customers that we seek to engage. Naturally, platforms prefer to engage a large fraction of customers (mainstream customers), while potentially sacrificing the engagement of users with niche preferences (tail customers). Thus, we introduce an alternative performance metric: for any policy π, let the set of *satisfied* customers (*i.e.,* customer preference vector realizations for which the policy achieves sublinear regret) be

$$U^\pi(\rho, p, T) := \{u \in \mathbb{R}^d : \mathcal{R}(T, \rho, p, u_0) = \mathcal{O}(T^\nu) \text{ for some } \nu \in [0, 1)\}\,. \tag{3}$$

Then, we define the Fraction of Satisfied Customers (FSC) under the customer distribution $\mathcal{P}$ as

$$FSC^\pi(\rho, p, T) = \mathbb{P}_{u_0 \sim \mathcal{P}}\left(U^\pi(\rho, p, T)\right)\,. \tag{4}$$

We will use the FSC metric to compare the performance of various policies under disengagement.

3.3. Classical Approaches

We may hope that widely-used approaches for product recommendations perform well in terms of the FSC metric defined in Eq. (4). Our next result considers the FSC of the class of *consistent* bandit learning algorithms Π^C (Definition 4 in Appendix D, based on Lattimore and Szepesvari 2016). This class 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 algorithms that balance an exploration-exploitation tradeoff. We also consider a simple greedy Bayesian updating policy (stated formally as Algorithm 3 in Appendix D), which greedily recommends the best estimated product and updates its posterior estimates (relative to the prior $\mathcal{P}$) based on observed customer feedback.

Proposition 2 **(Failure of Bandits and Greedy)**. *Let disengagement be salient for every customer: for every $u_0 \sim \mathcal{P}$, there is at least one product that may cause the customer to disengage, i.e., $|S(u_0, \rho)| < n$. Then, there exists a product set $\{V_i\}_{i=1}^n$ such that consistent bandit algorithms $\pi \in \Pi^C$ and the greedy Bayesian updating algorithm keep zero customers satisfied as $T \to \infty$, i.e.,*

$$\sup_{\pi \in \Pi^C} \inf_{\{V_i\}_{i=1}^n} FSC^\pi(\rho, p, T) = 0 \quad \textit{and} \quad \inf_{\{V_i\}_{i=1}^n} FSC^{GBU}(\rho, p, T) = 0\,.$$

The proof is given in Appendix B.1. Proposition 2 shows that consistent bandit and greedy algorithms result in linear regret for *every* customer realization. The proof is based on a construction of products such that the set of tolerable products satisfies $|\mathcal{S}(u_0,\rho)| < d$ for every u_0. Clearly, exploring outside this set can lead to disengagement. However, one cannot statistically estimate the true customer latent attributes u_0 without sampling products outside of the set; thus, all consistent bandit algorithms will sample outside the set $\mathcal{S}(u_0,\rho)$ infinitely many times (as $T \to \infty$), leading to customer disengagement 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). The design of bandit learning strategies fundamentally relies on the assumption that the time horizon T is exogenous, making exploration inexpensive. While intuition may suggest that greedy algorithms avoid over-exploration, they still involve *natural exploration* due to the noise in customer feedback (see, *e.g.,* Bastani et al. 2021), which may again cause the algorithm to over-explore and choose irrelevant products.

These results illustrate that there is a need 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 attributes 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.

3.4. Alternative Disengagement Models

Thus far, we presented the simplest possible disengagement model. However, our approach easily extends to alternative, more complex models of disengagement, *e.g.,*

1. The disengagement probability p may not be constant. It could depend on the time step t (capturing the customer's loyalty over time to the platform), or on the utilities derived from the recommendations thus far $\{\mu(u_0^\top a_i)\}_{i=1}^t$ (one poor recommendation may be less likely to cause disengagement if past recommendations were relevant). Then, we can express the customer's disengagement decision as:

$$\Pr[\Upsilon_t = 1 \mid a_t] = \begin{cases} 0 & \text{if } \; u_0^\top a_t \;\geq\; \rho\,, \\ p(t, u_0, a_1, \ldots a_t) & \text{otherwise.} \end{cases}$$

All results from our base model still hold as long as disengagement still occurs outside the set of tolerable products with some minimum positive probability, *i.e.,* $p(t,u_0,a_1,\ldots a_t) \geq \tilde{c} > 0$ for all $t, u_0, \{a_i\}_{i=1}^t$.

2. The customer disengagement decision might be *temporary*, *i.e.,* customers may decide to leave the platform for some length of time before returning to the platform again. Then, we abuse notation to let Υ_t denote the *total* time that the customer is disengaged due to all recommendations made until time t:

$$\Upsilon_t \mid a_t = \begin{cases} 0 & \text{if } \; u_0^\top a_t \;\geq\; \rho\,, \\ T^\delta & \text{otherwise}\,, \end{cases}$$

for some $\delta \leq 1$. Our previous models imposed $\delta = 1$ (the customer does not return for the remaining time horizon), while $\delta = 0$ models the classical bandit setting with no disengagement. In Appendix C.4, we show that our hardness result no longer holds when disengagement is sufficiently temporary (*i.e.,* when $\delta \leq 1/2$), but even in this setting, constraining exploration still yields significant empirical value.

4. Constraining Exploration

We have so far established that classical approaches fail on the product recommendation problem with customer disengagement. To gain an understanding of good policies in this setting, we first analyze a simplified scalar instance of our problem, reducing it to a known scheduling problem that has an optimal index-based policy. We analyze the structural properties of this policy, and show that it avoids arms that are likely sub-optimal for the entire time horizon. Motivated by this result, we propose the Constrained Bandit algorithm, which modifies standard bandit strategies by constraining the product set upfront using a novel integer programming formulation. The integer program leverages information on other customers on the platform to select a subset of products that are likely to be relevant for the incoming customer. Unlike classical approaches, the Constrained Bandit guarantees good performance on a significant fraction of customers.

4.1. Optimal policy for scalar case

First, consider a simplified version of our problem where customer response is a Bernoulli random variable, and each product's utility is independent of the utilities for other products, *i.e.,* $V_i = e_i, \forall i = 1, .., n$. We can cast this as a Markov Decision Process (MDP) with:

1. *Action space.* The set of products $\mathcal{A} = \{1, .., n\}$.
2. *Rewards.* If the customer is not disengaged, the reward for product recommendation a_t is a Bernoulli random variable with success probability $\theta_{a_t} := 1/(1 + \exp(-u_0^\top a_t))$. If the customer is disengaged, the reward is 0. The prior on the mean reward for product i is given by $\text{Beta}(\alpha_i, \beta_i)$.
3. *State space.* The state space $\mathcal{S}$ consists of $|\mathcal{A}|$ tuples, each tuple containing sufficient statistics for the utility distribution of product i. Let $F_t(i)$ denote the total number of times product i is recommended until time t, and $K_t(i)$ denote the total number of successes until time t. At time t,

$$s_t = \Big[(K_t(1), F_t(1)), .., (K_t(n), F_t(n)) \Big].$$

Let $s_t(i)$ denote the state space tuple associated with product i at time t, and let $\bar{q}_t(i)$ be the current estimate of the success probability of product i based on $s_t(i)$.

4. *Transition probabilities.* We have stochastic transition probabilities $\mathcal{T} : s \times \mathcal{A} \to s$ given by

$$\mathbb{P}\left(s_{t+1}(a_t) = s_t(a_t) + (1,1) \mid s_t, a_t\right) = \theta_{a_t},$$

$$\mathbb{P}\left(s_{t+1}(a_t) = s_t(a_t) + (0,1) \mid s_t, a_t\right) = 1 - \theta_{a_t}\,,$$

where only the state of a_t changes at time t. Note that the transition probabilities are also unknown but can be estimated using the current state s_t.

The objective is to maximize time-discounted expected reward given by:

$$\max_{a_1,a_2,..} \mathbb{E}\left[\sum_{t=0}^{\infty} \eta^t Y_{a_t}(S_t) \prod_{j=1}^{t} \Pr[\Upsilon_j = 0 \mid a_j]\right], \tag{CD}$$

where $\eta > 0$ is the discount factor. By Bellman's principle, the optimal policy solves the following recursive equation:

$$V^*(s) = \max_{a=\{1,..,k\}} \mathbb{E}\Big[\Pr[\Upsilon = 0 \mid a]\Big(Y_a(S) + \eta V^*(\hat{S}) | (S = s, a)\Big)\Big].$$

We will now map this recommendation problem to the "gold miner" machine scheduling problem, and use the celebrated Gittins Index Theorem to analyze the optimal policy.

Equivalence with the gold miner scheduling problem (Gittins et al. 2011): Consider the problem of extracting gold from n mines using a single machine. On a given day t, if the machine is used at mine a_t, then a total of $Y_{a_t}(S_t)$ units of gold can be extracted, where S_t is the state of the system at time t. However, the machine may break down forever with probability $1 - P_{a_t}(S_t)$ if mine a_t is selected in period t. The objective of the miner is to maximize the total (time-discounted) gold extracted by optimizing where to use the machine. In particular, we wish to solve:

$$\max_{a_1,a_2,..} \sum_{t=0}^{\infty} \mathbb{E}\left[\eta^t Y_{a_t}(S_t) \prod_{j=1}^{t} P_{a_j}(S_j)\right]. \tag{GM}$$

Lemma 1 **(§3.5 of Gittins et al. 2011)**. *The optimal policy of the gold miner's problem (GM) is an index policy. In particular, the index of arm i is given by:*

$$\nu_i(s) = \max_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{j=0}^{\tau} Y_i(S_j) \exp\Big(-\sum_{k=0}^{j} T_k(S_k, i)\Big) \middle| S_0 = s\right]}{\mathbb{E}\left[1 - \exp\left(-\sum_{k=0}^{\tau} T_k(S_k, i)\right) \middle| S_0 = s\right]},$$

where $T_k(S_k, i) = -\log(\eta^{1+\log_\eta(P_i(S_k))})$.

We now take Y_{a_t} to be a Bernoulli random variable with success probability θ_i if $a_t = i$. Since θ_i is unknown to the miner, she forms a belief of the mine's reward using the prior Beta(α_i,β_i). If we then take the probability of the machine not failing to be:

$$P_{a_t}(S_t) := 1 - \Pr[\Upsilon_t = 1 \mid a_t] = \begin{cases} \tilde{p}, & \text{if } \theta_{a_t} \leq \rho, \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Thus, we observe that problem (CD) and (GM) are identical, with $\tilde{p} := 1 - p$. Furthermore, Lemma 1 implies that the optimal policy of (GM) can be reduced to analyzing the state tuple $s_t(i)$ separately for each product i to compute its Gittins index. We will next prove bounds on the indices of each product in terms of the following stopping times:

$$\underline{\tau}_i^*(s) = \min\{t : \bar{q}_t(i) \leq \rho | S_0(i) = s\} \quad \text{and} \quad \bar{\tau}_i^*(s) = \min\{t : \bar{q}_t(i) > \rho | S_0(i) = s\},$$

where we recall that $\bar{q}_t(i)$ is the current estimate of the success probability of product i. Here, $\underline{\tau}_i^*(s)$ is a stopping time for the Markov process associated with product i, and denotes the first time at which there is a nonzero chance of the machine breaking down. Conversely, $\bar{\tau}_i^*(s)$ denotes the first time at which there is a zero chance of the machine breaking down.

Now, we define the state-dependent product sets

$$\begin{aligned}
\mathcal{K}_{opt}(s_t) &= \{i : \underline{\tau}_i^*(s_t) > 0\}, \\
\mathcal{K}_{sub}(s_t) &= \left\{ i : \bar{\tau}_i^*(s_t) > 0,\ \mathbb{E}\left[\tilde{p}^{\bar{\tau}_i^*(s_t)}\right] \leq \frac{\rho(1-\eta)}{\eta}\left((1-\tilde{p})\eta - \frac{\tilde{p}\eta}{1-\tilde{p}\eta}\right)\right\}.
\end{aligned}$$

Note that $|\mathcal{K}_{opt}(s_0)| > 0$ by our assumption that there is at least one product on the platform is tolerable the customer according to the Bayesian prior; similarly, $|\mathcal{K}_{sub}(s_0)| > 0$ by our assumption that at least one product is not tolerable to the customer according to the Bayesian prior. The next lemma (proof in Appendix B.2) relates the stopping times and product sets to the Gittins indices.

Lemma 2. *Consider the gold miner's problem (GM) with the probability of machine failure given by Eq.* (5)*. and also let*

$$\nu_{opt} = \min_{i \in \mathcal{K}_{opt}(s_t)} \nu_i(s_t) \quad \textit{and} \quad \nu_{sub} = \max_{i \in \mathcal{K}_{sub}(s_t)} \nu_i(s_t).$$

Then, $\nu_{opt} \geq \nu_{sub}$, and the optimal policy chooses a product in $\mathcal{K}_{opt}(s_t)$ and not in $\mathcal{K}_{sub}(s_t)$.

Lemma 2 shows that at time t, products in $\mathcal{K}_{sub}$ are ignored in favor of products in $\mathcal{K}_{opt}$. The next theorem shows that this ordering holds for the entire time horizon T with high probability.

Theorem 1. *Consider the gold miner's problem (GM) with the probability of machine failure given by Eq.* (5)*. Let $\theta_{opt}^* := \min_{i \in \mathcal{K}_{opt}(s_0)} \theta_i \geq \rho + \sqrt{\log_e(2)}$. Then, the optimal policy never selects products from $\mathcal{K}_{sub}(s_0)$ with probability at least*

$$\frac{1 - 2\exp(-(\theta_{opt}^* - \rho)^2)}{1 - \exp(-(\theta_{opt}^* - \rho)^2)} > 0.$$

Theorem 1 (proof in Appendix B.2) shows that the optimal policy constrains exploration to an *initially determined* set $\mathcal{K}_{opt}(s_0)$ for the *entire* time horizon T with high probability. This result sharply departs from classical bandit policies that would explore *all* products in an early exploration phase, particularly since we are in a setting where customer feedback from one product does not inform customer feedback from another product. This motivates our proposed Constrained Bandit algorithm.

4.2. Algorithmic Strategy

Our results thus far 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, 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 2 in the next section. Thus, we compromise performance on tail customers to achieve good performance on mainstream customers.

To this end, we introduce a set diameter parameter γ in our integer program formulation. This parameter tunes 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 in §4.4 based on Theorem 2.

4.3. Constrained Bandit Algorithm

We seek to find a restricted set of products that cater to a large fraction of customers (measured with respect to $\mathcal{P}$), but are not too "far" from each other (to constrain exploration). The following notation captures the likelihood that product i is relevant to a randomly sampled new customer:

Definition 2. $\mathcal{C}_i(\rho)$ 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 the set of tolerable products for a customer with attributes u_0 (Definition 1).

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. Estimating $\mathcal{C}_i(\rho)$ 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.

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 the indicator function

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

Note that γ and ρ are related. When the customer is less tolerant (ρ is large), we will choose smaller values of the set diameter γ and vice-versa; we specify how to choose γ in §4.4.

The objective is to select a subset 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:

$$\begin{aligned}
\mathbf{OP}(\gamma) = \max_{\mathbf{x},\mathbf{z}} \ & \sum_{i=1}^{n} C_i(\rho) x_i && \text{(6a)} \\
\text{s.t. } & z_{ij} \leq x_i, \quad i = 1, \ldots, n, && \text{(6b)} \\
& z_{ij} \leq x_j, \quad j = 1, \ldots, n, && \text{(6c)} \\
& z_{ij} \geq x_i + x_j - 1, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, && \text{(6d)} \\
& z_{ij} \leq \phi_{ij}(\gamma), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n, && \text{(6e)} \\
& x_i \in \{0, 1\} \quad i = 1, \ldots, n. && \text{(6f)}
\end{aligned}$$

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 defines whether product i is included in the restricted set, and $z_{i,j}$ (defined through constraints (6b)–(6d)) is an indicator variable for whether both products i and j are included in the restricted set. Constraint (6e) ensures that only products that are "close" to each other are selected.

Solving $\mathbf{OP}(\gamma)$ results in a subset 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.

We now describe the Constrained Bandit algorithm. From the UCB literature, an optimistic estimate of the customer utility from product V is:

$$f(\hat{u}_t, V) = \begin{cases} \max_{u \in \mathcal{Q}_t(\hat{u}_t)} u^\top V & \text{if } \mu(x) = x\,, \\ \hat{u}_t^\top V + \left(\tilde{C}\sqrt{4d\log(t)\log(2dT)}\right)\|V\|_{(\bar{X}_t+\lambda I)^{-1}} & \text{otherwise}\,. \end{cases} \tag{7}$$

Note that we have a different expression for the case where the customer response is linear ($\mu(x) = x$); this is because we can take advantage of the following closed form of the uncertainty ellipsoid around our estimate of u_0 at time t in the linear case:

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

Algorithm 1 Constrained Bandit(λ,γ)

```
Step 1: Constrained Exploration:
Solve $\mathbf{OP}(\gamma)$ to obtain constrained product subset $\Xi$; then, recommend a random product $a_1 \in \Xi$.
Step 2: Bandit Learning:
for $t \in [T]$ do
    Observe customer utility, $Y_t = \mu\left(u_0^\top a_t\right) + \varepsilon_t$.
    Let $\hat{u}_t$ be the unique solution to $\sum_{k=1}^{t-1}\left(Y_k - \mu(a_k^\top \hat{u}_t)\right)a_k = 0$.
    Let $a_t = \arg\max_{\{i \in \Xi\}} f(\hat{u}_t, V_i)$, for $f$ defined in Eq. (7).
    Recommend product $a_t$ at time $t$ if the customer is still engaged. Stop if the customer disengages from the platform.
end for
```

Algorithm 1 is a two-step procedure. First, the action space is restricted to the product subset 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 preferences and match her with her preferred product through repeated interactions. We use the OFUL algorithm (Abbasi-Yadkori et al. 2011) if the link function is identity, and the GLM UCB algorithm (Filippi et al. 2010) for general link functions. There are two input parameters: λ (a standard regularization parameter) and γ (the set diameter). We discuss the selection of γ and the corresponding tradeoffs next.

4.4. Theoretical Guarantees

Lemma 3 shows that the FSC (defined in Eq. (4)) of the Constrained Bandit is strictly positive, even in the worst case over product sets. In particular, we can always match some subset of customers to their preferred products by constraining the action space upfront. Again, this is in contrast with bandit and greedy algorithms (Proposition 2), which can achieve zero FSC. The proof is given in Appendix B.3.

Lemma 3. *Let disengagement be salient for every customer: for every $u_0 \sim \mathcal{P}$, there is at least one product offering that may cause the customer to disengage, i.e., $|S(u_0, \rho)| < n$. Then, in the worst case over all allowable product sets $\{V_i\}_{i=1}^n$, there exists a set diameter threshold γ_0 such that $\forall \gamma < \gamma_0$,*

$$\inf_{\{V_i\}_{i=1}^n} FSC^{CB(\lambda,\gamma)}(\rho, p, T) > 0\,.$$

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, which can trivially be done by restricting our product set Ξ to a single product. By construction of $\mathbf{OP}(\gamma)$, this will be the most popular preferred product, so a positive fraction of customers find this product optimal. Since these customers are immediately matched to their preferred product, we incur zero regret on this subset of customers.

Echoing the insights from Theorem 1, Lemma 3 shows that there is nontrivial value in restricting the product set upfront, which cannot be obtained through classical approaches. However, it considers the degenerate case of constraining exploration to a single product, which is clearly too restrictive in practice, especially when customers are relatively tolerant (*i.e.,* ρ is small). Ideally, we would want insight into how much the product set should be constrained as a function of the customer's tolerance parameter.

To answer this question, we consider a fluid approximation of the product space. Since $\mathbf{OP}(\gamma)$ is complex, we consider a continuous product space $V = [-1,1]^d$ to cleanly demonstrate the key tradeoff in constraining exploration: a larger product set has a higher probability of containing customers' preferred product, but also a higher risk of disengagement. Furthermore, 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 2 shows that the Constrained Bandit achieves sublinear regret for a fraction of customers under this albeit stylized setting. More importantly, it yields insights into how to choose the set diameter γ as a function of the customer's tolerance parameter ρ. We first define the following constants:

$$C_1 := \frac{1-\sqrt{1-\gamma^2/4}}{\sigma}, \quad C_2 := \frac{d\rho}{\sigma(1-\gamma)}, \quad \tilde{C} := (d+1)Y_{max} + \frac{2\sqrt{3+2\log(1+2L^2/\lambda)}\kappa_\mu Y_{max}}{c_\mu},$$
$$\bar{d} = \sqrt{2\left(1-\sqrt{(1-\gamma^2/4)}\right)}, \quad \text{and} \quad s := \max\{1, L^2/\lambda\}.$$

THEOREM 2. *Let $\mathcal{P} = \mathcal{N}(\bar{u}, \frac{\sigma^2}{d^2} I_d)$ and $V = [-1,1]^d$. Then, there exists a set $\mathcal{W}$ of latent customer attribute realizations (with positive probability under $\mathcal{P}$), such that for all $u_0 \in \mathcal{W}$, the cumulative regret of the Constrained Bandit is*

$$\begin{aligned}\mathcal{R}^{CB}(T,\rho,p,u_0) &\leq \begin{cases} \tilde{C} d \log(sT)\sqrt{2T\log(2dT)} & \text{if } \mu(x) = x\,, \\ 5\sqrt{Td\log(\lambda+TL)}\left(\sqrt{\lambda}(\bar{d}+1) + \xi\sqrt{\log(T) + d\log(1+TL)}\right) & \text{otherwise}\,. \end{cases} \\ &= \tilde{\mathcal{O}}\left(\sqrt{T}\right).\end{aligned}$$

Computing a lower bound on the volume of $\mathcal{W}$, we obtain that for any $\gamma < 1$,

$$FSC^{CB(\lambda,\gamma)}(\rho,p,T) \geq \frac{(1-2d\exp(-C_1))\left(\sqrt{4+C_2^2}-C_2\right)\exp\left(-C_2^2/2\right)}{2\sqrt{2\pi}}\,.$$

The proof of Theorem 2 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 for generalized models (Filippi et al. 2010) to bound the regret of the Constrained Bandit algorithm for the customer subset $\mathcal{W}$.

Theorem 2 provides an explicit characterization of 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 our FSC. 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.

In Appendix C.1, we approximate the optimal set diameter (that maximizes our lower bound on FSC) using a single parameter optimization problem that can be solved using numerical optimization. It is instructive to consider the simpler setting where $\bar{u}_i = 1/\sqrt{d}$ for all i, and $\rho < 1/\sqrt{d}$ (more tolerant customers); in this case, we can compute the optimal set diameter exactly as

$$\gamma^* = 1 - \rho\sqrt{d}\,.$$

This expression yields some useful comparative statics: we should choose a smaller set diameter γ when customers are less tolerant (ρ is large) or when the rank d of the latent features is high. In practice, we can tune the set diameter through cross-validation.

Remark 2 (Re-optimizing the constrained set). Our algorithm uses a fixed exploration set Ξ for the entire horizon T. A natural alternative is to update this set dynamically, *i.e.,* update our posterior on the customer's preference vector U_0 using noisy customer feedback, and re-solve $\mathbf{OP}(\gamma)$ using this posterior after every batch of observations (B time steps). Note that this departs from the structure of the optimal policy in Theorem 1. Perhaps surprisingly, we provide analytical and numerical evidence that frequent re-optimization *reduces* engagement time. This is because the variance of the noise in customer feedback ε is often high (*e.g.,* when click likelihood is low). The resulting uncertainty can cause the posterior update on the customer's latent attributes (and therefore the downstream IP solution) to fluctuate significantly and recommend worse products. We formalize this argument in Proposition 3 of Appendix B.4, where we prove that dynamic updating of the product set does not necessarily lead to improved performance. The performance gap is exacerbated in our numerical experiments (see Figure 2 in §5.1) since the variance of the noise is also typically unknown, and must be estimated on the fly. Thus, in practice, we recommend either re-optimizing with a large batch size B, or fixing the constrained set.

5. Numerical Experiments

We now compare the empirical performance of the Constrained Bandit with state-of-the-art Thompson sampling (Chapelle and Li 2011, Russo and Van Roy 2014) and greedy Bayesian updating. We study both synthetic data (§5.1) and real movie recommendation data (§5.2).

Benchmarks: We compare our algorithm with (i) linear Thompson Sampling (Russo and Van Roy 2014) and (ii) greedy Bayesian updating (referred to as MLE).

Constrained Thompson Sampling (CTS): We consider a Thompson Sampling version of the Constrained Bandit algorithm (see Algorithm 2 below). Recall that our approach allows for any bandit strategy after obtaining a restricted product set based on our (algorithm-independent) integer program $\mathbf{OP}(\gamma)$. To ensure a fair comparison, 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 2 Constrained Thompson Sampling (λ,γ)

```
Step 1: Constrained Exploration:
Solve $\mathbf{OP}(\gamma)$ to get the constrained set of products to explore over, $\Xi$. 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 $a_t = \arg\max_{\{i \in \Xi\}} \mu(u(t)^\top V_i)$ if the customer is still engaged.
    Observe customer utility, $Y_t = \mu(U_0^\top a_t) + \varepsilon_t$, and update $\hat{u}_t$ to be the unique solution of $\sum_{k=1}^{t-1} (Y_k - \mu(a_k^\top \hat{u}_t))\, a_k = 0$
    Stop if the customer disengages from the platform.
end for
```

5.1. 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[2] to be 2. We generate product features in each dimension uniformly

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

between -1 and 1. Similarly, latent user attributes are generated from a multivariate normal with with mean $[1/\sqrt{2}, 1/\sqrt{2}]^\top \in \mathbb{R}^2$ and variance $I_2 \in \mathbb{R}^{2\times 2}$, where we recall that I_d is the d-dimensional identity matrix. 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 $\varepsilon \sim \mathcal{N}(0,5)$, the customer tolerance ρ to be generated from a truncated $\mathcal{N}(0,1)$ distribution, and the total horizon length $T = 100$. 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 $p \in \{1\%, 10\%, 50\%, 100\%\}$, 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. Note that, due to its asymptotic nature, we cannot compute the FSC metric in finite sample. However, engagement time within a finite horizon is a closely related proxy, *i.e.,* as we showed in our earlier analysis, customer engagement is necessary to achieve low cumulative regret.

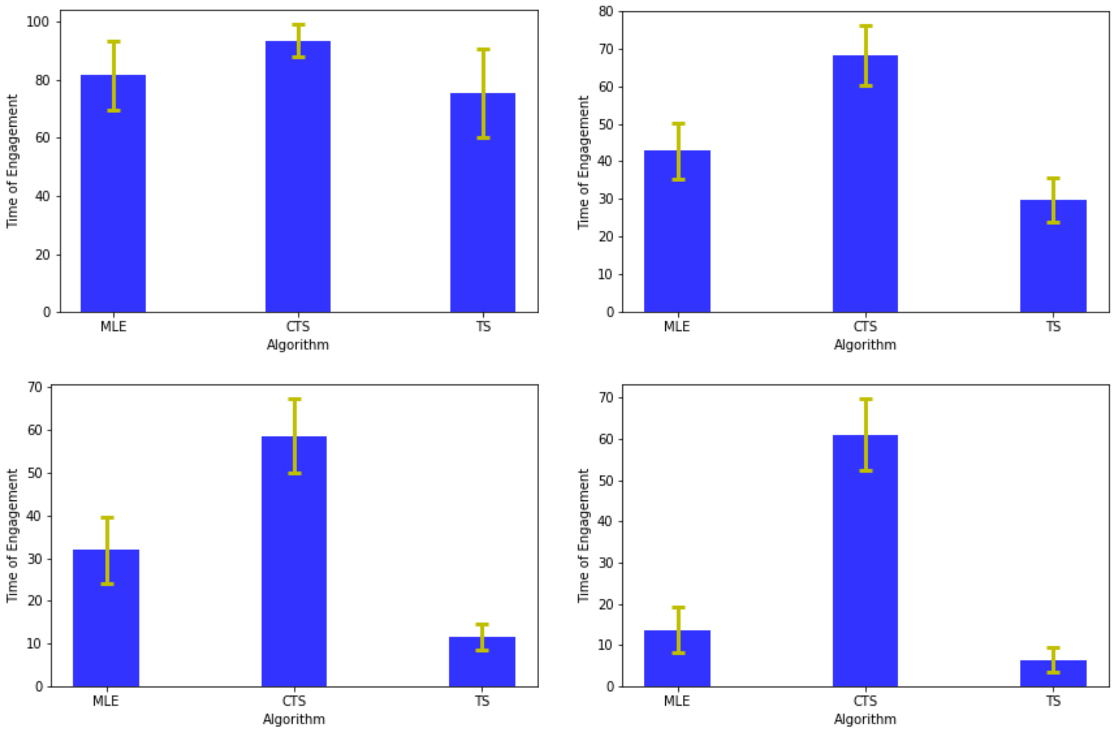

Figure 1 **Time of engagement and 90% confidence intervals averaged over 100 randomly generated customers for disengagement propensity p values of 1% and 10% (top row), 50%, and 100% (bottom row).**

Results: Figure 1 shows the customer engagement time averaged over 100 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 insignificant), TS performs well and CTS performs comparably. However, a greedy strategy is likely to converge to a suboptimal product outside of the customer's relevance set, causing the customer to eventually disengage. As we increase p, all algorithms perform worse, since customers become 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 when $p = 50\%$ and by a factor or 4.1 when $p = 100\%$. Thus, restricting the product set is critical when customer disengagement is significant.

Other Comparisons: Following Remark 2, we also compare the performance of an approach that re-optimizes the product set Ξ after every batch of observations (B time steps). In particular, after B interactions, we update our posterior mean and variance on the latent customer features, and resolve $\mathbf{OP}(\gamma)$ with the corresponding updated objective; we also exclude the product that caused the customer to disengage from our product set. In Figure 2, we plot the total time of engagement as a function of the batch size $B \in \{5, 10, 15\}$ (smaller B implies frequent re-optimization) and compare it with our approach and the other benchmarks. Perhaps surprisingly, we find that frequent re-optimization *reduces* engagement time. This is because frequent re-optimization can cause the posterior update on the customer's latent attributes (and therefore the downstream IP solution) to fluctuate significantly as a function of the idionsyncratic noise ε in the customer response, thereby recommending worse products; in contrast, selecting a static product set is robust to the (often large) noise in customer response. This intuition is formalized in Proposition 3 in Appendix B.4.

We also test the impact of modest misspecification of our key disengagement parameter ρ (see Appendix C.2), and find that the performance of our algorithm is robust to this uncertainty.

5.2. Case Study: Movie Recommendations

We now simulate CTS and the same benchmarks on a model calibrated with MovieLens, a publicly available movie recommendations dataset collected by GroupLens Research. This dataset is widely used in the academic community as a benchmark for recommendation and collaborative filtering algorithms (see Harper and Konstan 2016, for details). Importantly, we no longer have access to the true problem parameters (*e.g.,* ρ); we discuss simple heuristics for estimating these parameters.

5.2.1. 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

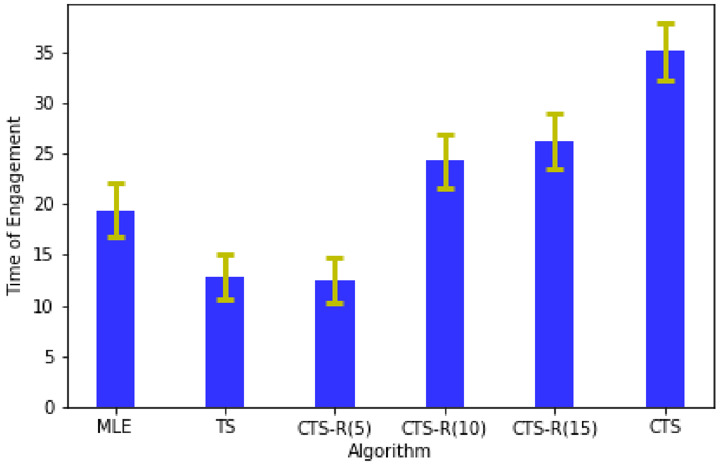

Figure 2 **Time of engagement and 95% CI (over 100 randomly generated customers) for $p = 10\%$, when the constrained set is reoptimized after every 5 (left), 10 (second from left), 15 (third from left) time periods or selected a-priori (right). Fixing a static constrained set outperforms dynamic updating.**

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. 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

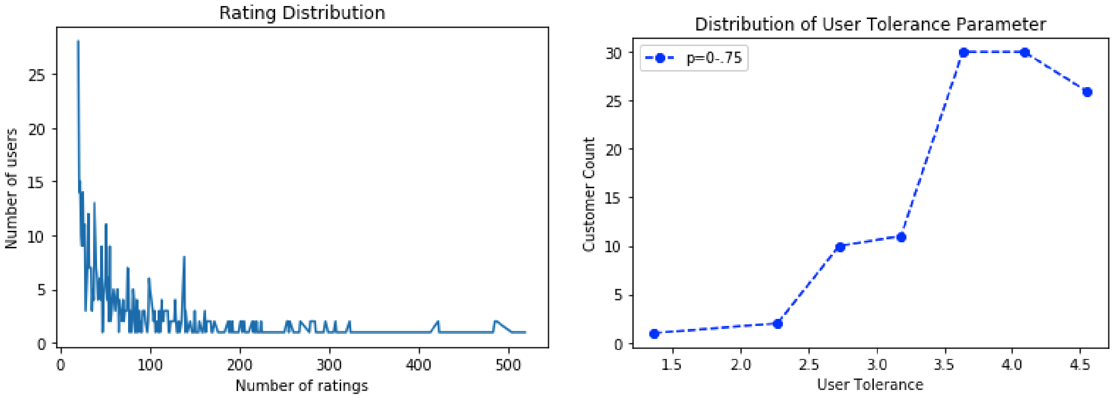

Figure 3 **On left, the histogram of user ratings in MovieLens data. On right, the empirical distribution of ρ, the customer-specific tolerance parameter, across all disengaged users for a fixed customer disengagement propensity $p = .75$. This distribution is robust to any choice of $p \in (0, .75]$**

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 than 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 relevance 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 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 minimum utility 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.

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 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 3, we also 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 more than 2, *i.e.,* they consider disengagement if the recommended movie's rating is less than 2 stars. As we may expect, very few disengaged users have a low estimated value of ρ, suggesting that they have high expectations on the relevance 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 does not change. Thus, we believe that this strategy is sound.

5.2.2. Results Similar to §5.1, 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 4 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 (§5.1). 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 1.8 when $p = 10\%$, by a factor of 2.14 when $p = 50\%$ and by a factor or 2.32 when $p = 100\%$. Thus, our main finding remains similar on this data: restricting the product set is critical when customer disengagement is significant.

6. Discussion and Conclusions

We study the classic problem of sequential product recommendations 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 relevance 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

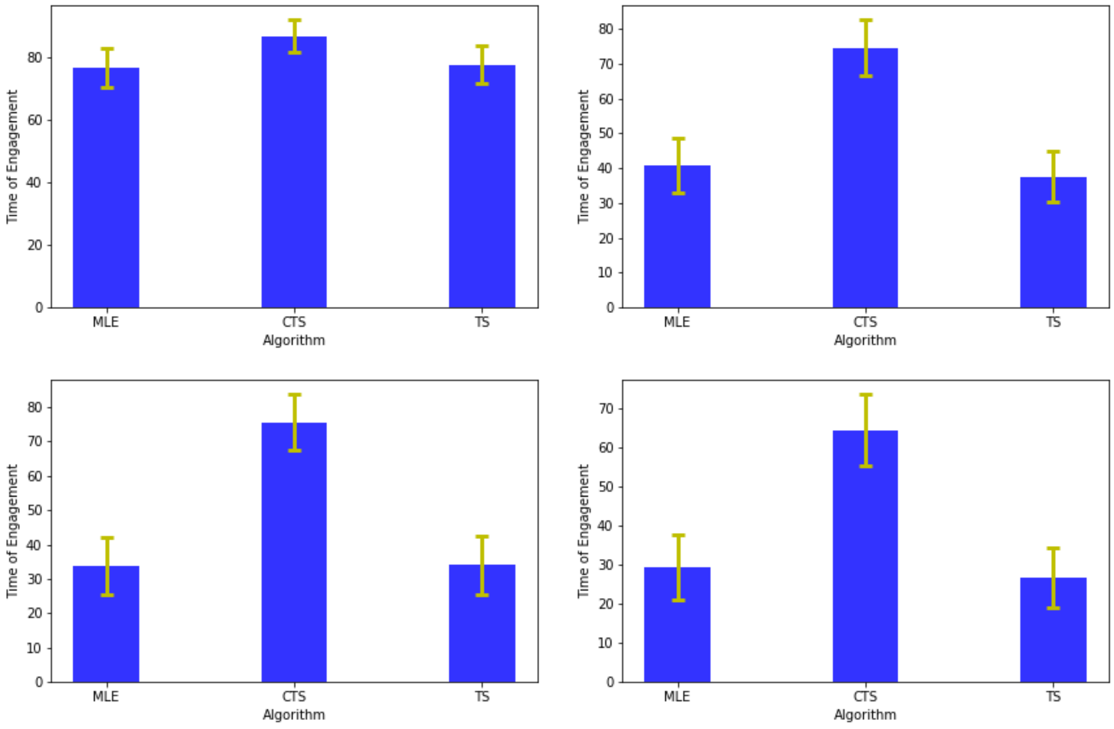

Figure 4 **Time of engagement and 95% confidence intervals on MovieLens data averaged over 1000 randomly generated customers for disengagement propensity p values of 1% (top left), 10% (top right), 50% (bottom left), and 100% (bottom right) respectively.**

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 show 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. Unfortunately, we find that classical bandit learning algorithms as well as simple greedy Bayesian updating perform poorly, and can fail to keep any customer engaged. Motivated by a reduction to a scheduling problem, we propose modifying bandit strategies by constraining the action space upfront using an integer program. We prove that this simple modification allows strong performance (*i.e.,* sublinear regret) for a significant fraction of customers. We also perform extensive numerical experiments on movie recommendations data that demonstrate the value of our approach towards improving customer engagement with the platform. Our results highlight a necessary tradeoff with clear managerial implications for platforms that seek to make personalized recommendations.

There are a number of practical considerations when deploying such an approach. First, our algorithm requires additional hyperparameters to calibrate the disengagement model, beyond the standard tuning parameters in bandit algorithms. We propose an approach to estimate these parameters using historical data (§5.2), and furthermore, show that modest misspecification does not significantly hurt performance (Appendix C.2). An alternative approach may be to jointly learn these parameters on the fly while making recommendations (see, for *e.g.,* Li et al. 2017). Second, we build on the classic collaborative filtering model, which does not have any side information. However, side information can be easily accommodated into the collaborative filtering framework, *e.g.,* through observed user/product features (Jain and Dhillon 2013), or tensors capturing multiple outcome variables (Farias and Li 2019). Such information, when available, can significantly speed up learning, thereby improving customer satisfaction and retention.

More broadly, our work leads to a number of interesting directions for future research. For instance, our disengagement model is based on customer utility thresholds. However, customer disengagement behaviour may not be homogeneous across the platform, may change over time, or may be dictated by external circumstances. Defining what constitutes customer disengagement, and developing empirical estimation techniques for identifying customer disengagement on platforms remains an important direction for future research. Moreover, we focus on the setting where the platform recommends a single product to the customer at each time step. Many platforms recommend assortments of products. This can help engage variety-seeking customers (Kahn 1995), as well as communicate more information about the platform's offerings and the customer's preferences, *e.g.,* in some cases, the assortment drives the customer's opinion on whether to engage with the platform (Ferreira et al. 2019). Thus, an interesting direction is to design algorithms that offer customers subsets of products to infer the underlying customer choice model (Chen et al. 2018, Feng et al. 2018) while accounting for customer disengagement. Finally, while we provide a numerical study calibrated on real data, a field experiment could shed more light into the value of our approach in practice.

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. 2021. Predicting with proxies: Transfer learning in high dimension. *Management Science* **67**(5) 2964–2984.

Bastani, Hamsa, Mohsen Bayati, Khashayar Khosravi. 2021. Mostly exploration-free algorithms for contextual bandits. *Management Science* **67**(3) 1329–1349.

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, Xi, Yuanzhi Li, Jieming Mao. 2018. A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2504–2522.

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.

Duembgen, L. 2010. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063* .

Farias, Vivek F, Andrew A Li. 2019. Learning preferences with side information. *Management Science* **65**(7) 3131–3149.

Feng, Yifan, Rene Caldentey, T. Christopher Ryan. 2018. Robust learning of consumer preferences. *Available at SSRN 3215614* .

Ferreira, Kris, Sunanda Parthasarathy, Shreyas Sekar. 2019. Learning to rank an assortment of products. *Available at SSRN 3395992* .

Filippi, Sarah, Olivier Cappe, Aurélien Garivier, Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*. 586–594.

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

Garivier, Aurélien, Emilie Kaufmann. 2016. Optimal best arm identification with fixed confidence. *Conference on Learning Theory*. PMLR, 998–1027.

Gittins, J, K Glazebrook, R Weber. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

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.

Jain, Prateek, Inderjit S Dhillon. 2013. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626* .

Jain, Prateek, Praneeth Netrapalli, Sujay Sanghavi. 2013. Low-rank matrix completion using alternating minimization. *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 665–674.

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* .

Kahn, Barbara E. 1995. Consumer variety-seeking among goods and services: An integrative review. *Journal of retailing and consumer services* **2**(3) 139–148.

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. *Columbia Business School Research Paper* (18-52).

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, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, Ameet Talwalkar. 2017. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research* **18**(1) 6765–6816.

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

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.

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

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. *arXiv preprint arXiv:1802.05693* .

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.

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

Appendix

A. Airline Campaign Details

Appendix A.1 overviews how we trained the personalized relevance score in §2, and Appendix A.2 argues why omitted variables do not bias our estimated treatment effects.

A.1. Collaborative filtering

Recall from our preliminaries in §3 that we wish to recover matrix $A = U^\top V$. Here, we interpret $U_i^\top V_j$ as the utility of customer $i \in [m]$ for product $j \in [n]$. Given a set $\mathcal{I}$ of observed utilities (noisy entries of A), the objective is to recover estimates of $U \in \mathbb{R}^{d\times m}$ and $V \in \mathbb{R}^{d\times n}$ so that we can infer unobserved entries of A (to make new recommendations). This can be written as the following error minimization problem

$$\min_{U,V} \sum_{(i,j)\in\mathcal{I}} (A_{ij} - U_i^\top V_j)^2 \,. \tag{8}$$

Problem (8) is typically non-convex due the bilinearity of the objective function. Alternating minimization — where we iteratively fix either U or V and optimize over the other quantity — is a popular approach because of its ease of implementation, strong performance, and theoretical guarantees. In particular, when the observed entries of A are distributed uniformly at random, alternating minimization recovers A with $\mathcal{O}(nd^{4.5}\log(n))$ observations (Jain et al. 2013).

We use a Python implementation[3] of alternating minimization for our collaborative filtering estimation throughout the paper. Following standard practice, we tune the rank of the underlying model on a held-out validation set. Although we do not have access to additional information on customers (beyond their previous clicks, that we take to be the noisy entries of A), such information can be easily accommodated in collaborative filtering. For instance, one can incorporate user/product features (see, for e.g., Jain and Dhillon 2013), or multiple outcome variables via tensors (see, for e.g. Farias and Li 2019).

A.2. Omitted variable bias

We do not control for a number of customer-specific attributes (e.g., customer loyalty tier/status, unobserved travel preferences) due to data limitations. To assuage concerns about omitted variable bias, we now provide an econometric argument that customer-specific omitted variables will *not* bias our treatment effect estimate.

For simplicity, consider a linear model given by

$$y_i = \beta_1 t_i + \beta_2 x_i + \epsilon_i \,,$$

where y_i is the dependent variable, t_i is the treatment, x_i is an omitted explanatory variable, ϵ_i is idiosyncratic noise, and the scalar β_1 is the treatment effect of interest. When x_i is missing from the regression specification, the OLS estimator of β_1 is given by

$$\frac{\partial}{\partial t}E[y|t] = \frac{\partial}{\partial t}E[\beta_1 t + \beta_2 x + \epsilon_i|t] = \frac{\partial}{\partial t}\left[\beta_1 t + E[\beta_2 x|t]\right] = \beta_1 + \beta_2 \frac{\partial}{\partial t}E[x|t] \,.$$

As expected, if the omitted variable x is uncorrelated with the outcome y, then $\beta_2 = 0$ and we obtain an unbiased estimate of the treatment effect β_1. But, in practice, β_2 is unlikely to be 0 since customer-specific

[3] `https://spark.apache.org/docs/latest/ml-collaborative-filtering.html`

preferences likely affect the probability of a click outcome. Instead, we argue that the omitted variable x is uncorrelated with the treatment t, in which case $\frac{\partial}{\partial t}E[x|t]=0$ so we still obtain an unbiased estimate of the treatment effect β_1 even if $\beta_2 \neq 0$.

The airline decided the destinations that they will promote *before* the start of the first campaign. Thus, there is no correlation between customer-specific preferences (such as loyalty tier/status or travel history) with the recommended destinations, *i.e.,* the relevance score of the 6^{th} campaign for a customer is uncorrelated with unobserved customer-specific preferences. Hence, while these omitted variables are likely predictive of the customer response, they do not bias the estimated treatment effect.

B. Proofs of Main Results

This section provides the proofs for all results in the paper.

B.1. Lower bounds

First, we prove hardness by constructing a simple instance where linear regret is unavoidable.

Proof of Proposition 1: Consider WLOG the case when $d = 2$. Then, $u_0 \sim \mathcal{N}(0, \sigma^2 I_2)$. Furthermore, $V_1 = [1,0]$ and $V_2 = [0,1]$. Clearly, Product 1 is optimal when $u_{0_1} > u_{0_2}$ and vice versa. For any ρ, consider the following events: $\mathcal{E}_1 = \{u_{0_1} < \rho < u_{0_2}\}$, and $\mathcal{E}_2 = \{u_{0_2} < \rho < u_{0_1}\}$. Then on $\mathcal{E}_1$, recommending product 1 leads to customer disengagement with probability p and on $\mathcal{E}_2$, recommending product 2 leads to customer disengagement with probability p. But, $C := \mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\mathcal{E}_2) = \mathbb{P}(Z < \rho/\sigma)(1 - \mathbb{P}(Z < \rho/\sigma)) > 0$, where Z denotes the standard normal random variable and the last inequality follows since ρ is finite by assumption. Any policy π has two options at time 1: either to recommend product 1 or to recommend product 2. First consider the case when $a_1 = 1$ and notice that

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

Similarly, when $a_1 = 2$, $\mathbb{E}_{u_0\sim\mathcal{P}}\left[\mathcal{R}^\pi(T,\rho,p,u_0)\right] \geq CpT$. Hence,

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

The proof follows similarly for any $d > 2$ since the probability of disengagement continues to be strictly positive in the initial round. □

Before we prove Proposition 2, we prove a 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 4. *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 of Lemma 4: The proof strategy is similar to that of Theorem 1 in Lattimore and Szepesvari (2016) with two main steps. In Step 1, we show that $\limsup_{t\to\infty}\log(t)\|V-V_*\|^2_{X_t^{-1}} \leq \frac{\Delta_V^2}{2(1-\nu)}$. Then, in Step 2, we connect this result to the matrix norm on the features of V which leads to the final result. We skip the details for the sake of brevity and refer the interested readers to Lattimore and Szepesvari (2016). □

Proof of Proposition 2: Part 1 (Bandit Failure): Since, this result is over all possible product feature settings, we consider the following setting: Let μ be the identity function (linear case), and there be d total products in $\mathbb{R}^d$, with latent product features $V_i = e_i$, the i^{th} basis vector. By assumption $|S(u_0,\rho)| < d$. We will show that 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 increase ρ and make it sufficiently large 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.

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 the simplified setting, there are d products in total $(n = d)$ and the feature vector of the i^{th} product is the i^{th} basis vector. 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 (1)). 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 $T_j(t) = \mathbb{E}\left[\sum_{f=1}^t \mathbb{1}\{a_f^\pi = j\}\right]$. $T_j(t)$ is the total number of times the j^{th} product is offered until time t under policy π. Next consider the following:

$$\begin{aligned}\text{lim sup}_{t\to\infty}\log(t)\|e_1\|^2_{X_t^{-1}} &= \text{lim sup}_{t\to\infty}\log(t)e_1^\top X_t^{-1}e_1 = \text{lim sup}_{t\to\infty}\log(t)e_1^\top\mathbb{E}\left[\sum_{f=1}^t a_f a_f^\top\right]^{-1} e_1 \\ &= \text{lim sup}_{t\to\infty}\log(t)e_1^\top\left[R_t\right]^{-1}e_1 = \text{lim sup}_{t\to\infty}\log(t)\left(\frac{1}{T_1(t)}\right) \\ &\geq \text{lim sup}_{t\to\infty}\log(t)\left(\frac{1}{T_1(\bar{t})}\right) = \infty\,. \end{aligned} \tag{9}$$

Where the second to last inequality follows from 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,

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

which implies that $\exists a_i$ in the action space such that the condition of Lemma 4 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 large and p is positive, customers are guaranteed to disengage from the platform eventually. This leads to a linear rate of regret for all customers. Hence,

$$\sup_{\pi\in\Pi^C}\inf_{\{V_i\}_{i=1}^n} FSC^\pi(\rho,p,T) = 0\,.$$

Part 2 (Greedy Failure): Recall, that 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. Let us assume, WLOG, that the GBU policy picks product 1 to recommend. We have two cases to analyze: (i) product 1 is sub optimal for the realized latent attribute vector, u_0, (ii) product 1 is optimal for the realized latent attribute vector, u_0. Let us consider case (i) when product 1 is suboptimal. In this case, if we let ρ to be large enough ($\rho > u_0^\top V_1$), so that the customer leaves with probability p in the current round. Hence, for all such customers

$$\mathcal{R}^\pi(T,\rho,p,u_0) \geq T \cdot p = pT\,.$$

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 outside of the relevance threshold. This would again lead to a 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, let G^t to be the event that the GBU policy switches to some other product from product 1 by time t. Then,

$$\mathbb{P}(G^t) = \mathbb{P}\left(\cup_{i=2..n} \cup_{j=1..t} (E_i^j)^c\right) \geq \mathbb{P}\left((E_i^j)^c\right)\,,\ \forall i = 2,..,n, \forall j = 1,..,t\,.$$

We will lower bound the probability of product 1 not being the optimal product for some 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 analyze the probability of switching from product 1 to product 2 after round 1 ($(E_2^1)^c$). First note that, $E_i^t = \{V_1^t \hat{u}_t - V_i^\top \hat{u}_t \geq 0\}$, which implies

$$(E_i^t)^c = \{V_i^t \hat{u}_t - V_1^\top \hat{u}_t > 0\} = \{(V_i - V_1)^\top (\hat{u}_t - u_0) > \Delta_i\}\,,$$

where $\Delta_i = V_1^\top u_0 - V_i^\top u_0$. Now, note that $\hat{u}_t = \left[\sum_{f=1}^t a_f a_f^\top + \frac{\xi^2}{\sigma^2} I_d\right]^{-1} [a_{1:t}]^\top Y_{f=1:t}$. Hence,

$$\hat{u}_1 = \begin{bmatrix} 1 + \frac{\xi^2}{\sigma^2} & 0 \\ 0 & \frac{\xi^2}{\sigma^2} \end{bmatrix}^{-1} \begin{bmatrix} Y_1 & 0 \\ 0 & 0 \end{bmatrix} = \left[\frac{\sigma^2 Y_1}{\sigma^2 + \xi^2}, 0\right]\,.$$

Therefore, we are interested in the event

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

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 $\mathbb{P}(G^t) \geq C_4$. Following the same regret argument as before, we have that for all such customers, $\mathcal{R}^{\text{GBU}}(T,\rho,p,u_0) = C_4 \cdot T$. The argument for $d > 2$ follows similarly since with positive probability, the GBU policy would either get stuck at a sub-optimal arm or would switch to a sub-optimal arm. Hence,

$$\sup_{\pi \in \Pi^C} \inf_{\{V_i\}_{i=1}^n} FSC^\pi(\rho,p,T) = 0\,.$$

□

B.2. Optimal policy for scalar case

Proof of Lemma 2: We will suppress the dependence of $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ on s_t for ease of notation in what follows. Assume WLOG that arm 1 belongs to $\mathcal{K}_{opt}$ and arm 2 belongs to $\mathcal{K}_{sub}$. We prove the result above by showing an upper bound on the index of arm 1 and a lower bound on the Gittin's index of arm 2. Consider the index of arm (1) and note that

$$\nu_1(s) = \max_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{j=0}^{\tau} Y_1(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,1)\right) \middle| S_0 = s\right]}{\mathbb{E}\left[1 - \exp\left(-\sum_{k=0}^{\tau} T_k(S_k,1)\right) \middle| S_0 = s\right]} \tag{10}$$

$$\geq \frac{\mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} Y_1(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,1)\right) \middle| S_0 = s\right]}{\mathbb{E}\left[1 - \exp\left(-\sum_{k=0}^{\underline{\tau}_1^*(s)} T_k(S_k,1)\right) \middle| S_0 = s\right]}. \tag{11}$$

Next, notice that for any $t < \underline{\tau}_1^*(s), \bar{q}_t(1) > \rho$. Furthermore, note that the expected gold mined for any $t \leq \underline{\tau}_1^*(s)$ is greater than ρ. In particular, for any $t \leq \underline{\tau}_1^*(s)$, $\mathbb{E}[Y_1(S_t)] = \bar{q}_t(1) > \rho$. Re-evaluating the numerator of the RHS in (10), we have that

$$\mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} Y_1(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,1)\right) \middle| S_0 = s\right] \geq \mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} \eta^j \rho\right] = \rho \mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} \eta^j\right].$$

Similarly, focusing on the numerator of the RHS in (2), we have that

$$\mathbb{E}\left[1 - \exp\left(-\sum_{k=0}^{\underline{\tau}_1^*(s)} T_k(S_k,1)\right) \middle| S_0 = s\right] = \mathbb{E}\left[1 - \prod_{j=1}^{\underline{\tau}_1^*(s)} \eta\right] = \mathbb{E}\left[1 - \eta^{\underline{\tau}_1^*(s)}\right].$$

Hence,

$$\nu_1(s) \geq \rho \frac{\mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} \eta^j\right]}{\mathbb{E}\left[1 - \eta^{\underline{\tau}_1^*(s)}\right]} \geq \rho \mathbb{E}\left[\sum_{j=0}^{\underline{\tau}_1^*(s)} \eta^j\right] \geq \rho\eta,$$

where the last inequality follows by assumption that $\underline{\tau}_1^*(s) > 0$. Note that the above analysis was independent of the arm's index. Hence, $\forall i \in \mathcal{K}_{opt},\ \nu_s \geq \rho\eta \implies \nu_{opt} \geq \rho\eta$, where the last inequality follows from the definition of ν_{opt}.

Next we consider the index of arm 2 and prove an upper bound on its value. First note by definition that

$$\nu_2(s) = \max_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{j=0}^{\tau} Y_2(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right) \middle| S_0 = s\right]}{\mathbb{E}\left[1 - \exp\left(-\sum_{k=0}^{\tau} T_k(S_k,2)\right) \middle| S_0 = s\right]}. \tag{12}$$

Let τ^* be the optimal index in the expression above. We will relate τ^* to $\bar{\tau}_2^*$. To show an upper bound, we will analyze two cases: (i) $\tau^* \geq \bar{\tau}_2^*$ or (ii) $\tau^* \leq \bar{\tau}_2^*$. We start by considering the case (i) when $\tau^* \geq \bar{\tau}_2^*$. Rewriting (12), we have that

$$\nu_2(s) = \frac{\mathbb{E}\left[\sum_{j=0}^{\bar{\tau}_2^*} Y_2(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right) + \sum_{j=\bar{\tau}_2^*+1}^{\tau^*} Y_2(S_j) \exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right) \middle| S_0 = s\right]}{\mathbb{E}\left[1 - \tilde{p}^{\bar{\tau}_2^*} \eta^{\tau^*} \middle| S_0 = s\right]}. \tag{13}$$

Let us consider the numerator of (13) and note that

$$\mathbb{E}\left[\sum_{j=0}^{\bar{\tau}_2^*} Y_2(S_j)\exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right)\right] \le \sum_{j=0}^{\bar{\tau}_2^*} \rho(\tilde{p}\eta)^j \le \rho\frac{\tilde{p}\eta}{1-\tilde{p}\eta}\,,$$

where the first inequality holds directly by the definition of $\bar{\tau}_2^*$ and $T_k(S_k,2)$. And the second inequality holds by evaluating the sum of the geometric series. Next, consider

$$\begin{aligned}\mathbb{E}\left[\sum_{j=\bar{\tau}_2^*+1}^{\tau^*} Y_2(S_j)\exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right)\right] &\le \mathbb{E}\left[\sum_{j=\bar{\tau}_2^*+1}^{\infty} Y_2(S_j)\exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right)\right]\\ &\le \mathbb{E}\left[\sum_{j=\bar{\tau}_2^*+1}^{\infty} \exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right)\right]\\ &= \mathbb{E}\left[\sum_{j=\bar{\tau}_2^*+1}^{\infty} \tilde{p}^{\bar{\tau}_2^*}\eta^j\right] \le \mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\frac{\eta}{1-\eta}\right]\,.\end{aligned}$$

where the first inequality follows because we are summing up positive numbers, second inequality follows because rewards are upper bounded by 1, and the last inequality follows by summing up the infinite geometric series. Hence, combining the above two bounds, we get an overall upper bound on the numerator of (12) by:

$$\mathbb{E}\left[\sum_{j=0}^{\tau^*} Y_2(S_j)\exp\left(-\sum_{k=0}^{j} T_k(S_k,2)\right)\middle| S_0=s\right] \le \frac{\rho\tilde{p}\eta}{(1-\tilde{p}\eta)} + \mathbb{E}\left[\frac{\tilde{p}^{\bar{\tau}_2^*}\eta}{(1-\eta)}\right]\,.$$

Finally, focusing on the denominator of (12), we have that $\mathbb{E}\left[1-\tilde{p}^{\bar{\tau}_2^*}\eta^{\tau^*}\right] \ge \mathbb{E}\left[1-\tilde{p}\right] = 1-\tilde{p}$. Combining the upper bound on the numerator and the lower bound on the denominator, we get that

$$\nu_2(s) \le \frac{1}{1-\tilde{p}}\left(\frac{\rho\tilde{p}\eta}{(1-\tilde{p}\eta)} + \mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\frac{\eta}{1-\eta}\right]\right)\,.$$

Note that so far we have only used the condition that $\bar{\tau}_2^* > 0$. Since all arms $j \in \mathcal{K}_{sub}$ also satisfy this assumption, we have that $\forall j \in \mathcal{K}_{sub}$,

$$\nu_j(s) \le \frac{1}{1-\tilde{p}}\left(\frac{\rho\tilde{p}\eta}{(1-\tilde{p}\eta)} + \mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\frac{\eta}{1-\eta}\right]\right) \implies \nu_{sub} \le \frac{1}{1-\tilde{p}}\left(\frac{\rho\tilde{p}\eta}{(1-\tilde{p}\eta)} + \mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\frac{\eta}{1-\eta}\right]\right)\,.$$

Finally, recall by assumption that arms belonging to $\mathcal{K}_{sub}$ also satisfy

$$\mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\right] \le \frac{\rho(1-\eta)}{\eta}\left(p\eta - \frac{\tilde{p}\eta}{1-\tilde{p}\eta}\right)\,.$$

Rearranging, we get that

$$\nu_{sub} \le \frac{1}{1-\tilde{p}}\left(\rho\frac{\tilde{p}\eta}{1-\tilde{p}\eta} + \mathbb{E}\left[\tilde{p}^{\bar{\tau}_2^*}\frac{\eta}{1-\eta}\right]\right) \le \rho\eta \le \nu_{opt}\,.$$

Hence, we have shown that $\nu_{sub}(s) \le \nu_{opt}(s)$. Since the optimal policy is an index policy (by Lemma 1), we have that arms in $\mathcal{K}_{opt}$ are preferred over arms in in $\mathcal{K}_{sub}$, proving the final result. The case of $\tau^* < \bar{\tau}_2^*(s)$ follows similarly, since by assumption $\bar{\tau}_2^*(s) > 0$. We skip the details for the sake of brevity.

Proof of Theorem 1: Assume WLOG that $|\mathcal{K}_{opt}(s_0)| = 1$. We will show that the single arm in $\mathcal{K}_{opt}(s_0)$ is always preferred over other arms in $\mathcal{K}_{sub}(s_0)$. The result would follow without loss of generality since at any time if there are more than one arms in $\mathcal{K}_{opt}(s_0)$, they are all preferred over arms in $\mathcal{K}_{sub}(s_0)$. Furthermore, no arm can leave $\mathcal{K}_{opt}$ without being pulled at least once. Hence, even if the size of arms in $|\mathcal{K}_{opt}(s_0)| > 1$, either arms start to drop off from the set and eventually the set contains only a singe arm, or they continue staying in the set over time. In either case, arms in $\mathcal{K}_{sub}(s_0)$ would be eventually compared to arms in $\mathcal{K}_{opt}(s_0)$ before being pulled. Hence, in what follows we will assume that the first arm is contained in $\mathcal{K}_{opt}(s_0)$.

Recall that the initial state, and Lemma 2 ensures that arm 1 will be chosen in the first time period. Hence, after the first pull, only the state space of the first arm changes and not of any other arm. Next, notice again by Lemma 1, that if $S_2(1)$ is such that $\underline{\tau}_1^*(S_2(1)) > 0$, then we can use Lemma 2 to show that arm 1 will still be preferred over arms in $\mathcal{K}_{sub}(S_2)$ at time 2. This follows because the arm's state for all arms in $\mathcal{K}_{sub}(s_0)$ have not changed so far. This argument continues to hold for all time t. Hence, if a switch from arm 1 to another arm in $\mathcal{K}_{sub}(S_t)$ happens, arm 1 must leave $\mathcal{K}_{opt}(S_t)$ at such time t. Conversely, if $\underline{\tau}_1^*(S_t) > 0,\ \forall t = 1, ..\infty$, then arm 1 never leaves $\mathcal{K}_{opt}(S_t)$ and is always the preferred arm. We will show a lower bound on the probability of this event. First recall by definition that

$$\underline{\tau}_1^*(s_t) = \min\{t : \bar{q}_t(1) \leq \rho | S_0(i) = s\},$$

denotes the first time when the estimated probability of success of arm 1 drops below the threshold ρ, after starting in state s. Then, trivially $\underline{\tau}_1^*(s_t)$ equals 0 when $\bar{q}_t(1) \leq \rho$. But at any time t, $\bar{q}_t(1) = \frac{\alpha_1 + K_t}{\alpha_1 + \beta_1 + F_t}$. Here, K_t and F_t denote the total successes and total pulls of arm 1. Hence, at time t, if

$$\alpha_1 + K_t/\alpha_1 + \beta_1 + F_t \leq \rho \implies K_t/F_t - \theta_1 \leq \rho + \rho/F_t\,(\alpha_1 + \beta_1) - \alpha_1/F_t - \theta_1\,.$$

Hence, at any time t, the estimated probability of success to be below threshold ρ is given by:

$$\mathbb{P}\left(K_t/F_t - \theta_1 \leq (\rho - \theta_1) + \alpha_1/F_t\,(\rho - 1) + \beta_1/F_t\right)\,.$$

Let $\beta_1 = \bar{\kappa}\alpha_1$, for some $\bar{\kappa} < 1$ and assume that $1 - \rho - \bar{\kappa} > 0$. Then,

$$\mathbb{P}\left(K_t/F_t - \theta_1 \leq (\rho - \theta_1) + \alpha_1/F_t\,(\rho + \bar{\kappa} - 1)\right) \leq \mathbb{P}\left(\theta_1 - K_t/F_t \geq (\theta_1 - \rho)\right) \leq \exp\left(-F_t(\theta_1 - \rho)^2\right)\,, \tag{14}$$

where the last inequality follows by a direct application of Hoeffding's inequality for bounded random variables. Let $\mathcal{E}_t = \{\bar{q}_t(1) > \rho\}$. Then we are interested in lower bounding the probability of $A := \cap_{t=1}^{\infty}\mathcal{E}_t$. But

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}\left((\cap_{t=1}^{\infty}\mathcal{E}_t)^c\right) = 1 - \mathbb{P}\left(\cup_{t=1}^{\infty}(\mathcal{E}_t)^c\right) \\
&\geq 1 - \sum_{t=1}^{\infty}\mathbb{P}((\mathcal{E}_t)^c) \geq 1 - \sum_{t=1}^{\infty}\exp\left(-F_t(\theta_1 - \rho)^2\right) \\
&\geq 1 - \sum_{t=1}^{\infty}\exp\left(-t(\theta_1 - \rho)^2\right) \\
&= 1 - \frac{\exp(-(\theta_1 - \rho)^2)}{1 - \exp(-(\theta_1 - \rho)^2)} = \frac{1 - 2\exp(-(\theta_1 - \rho)^2)}{1 - \exp(-(\theta_1 - \rho)^2)}\,,
\end{aligned}$$

where the first inequality follows using union bound, the second inequality follows by (14) and the second last equality follows by using the geometric sum of infinite series. This proves the final result since $\theta_1 \geq \theta_{opt}^*$.

□

B.3. Proofs for Constrained Bandit

First, we prove Lemma 3, which shows that the FSC of the Constrained Bandit is strictly positive, even in the worst case over all product sets.

Proof of Lemma 3: Consider any feasible $\rho > 0$ and let γ_0 be the maximum constraining parameter such that only a single product remains in the constrained exploration set. Than for any $\gamma \leq \gamma_0$ $\mathbf{OP}(\gamma)$ picks a single product $(\tilde{i})$ in the exploration phase. Now for any such γ consider

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

Then we have that $\forall u_0 \in \mathcal{W}_{\lambda,\gamma}$, 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 $\mathbb{P}(\mathcal{W}_{\lambda,\gamma}) > 0$. For example, if 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 in fact 0 since the platform recommends their optimal product. Therefore, for any ρ, $FSC^{CB(\lambda,\gamma)}(\rho,p,T) > 0$, which proves the final result. □

Next, we prove Theorem 2, upper bounding the regret among some (mainstream) customers. We begin by defining $L_{t,\rho,p}$, an indicator that captures whether the customer is still engaged at time t:

Definition 3. Let,

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

Clearly, $\mathbb{1}\{L_{T,\rho,p} = 1\} = \Pi_{t=1}^{T} \mathbb{1}\{\Upsilon_t = 0\}$, where we recall that Υ_t is the disengagement decision of the customer at time t. 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.

Proof of Theorem 2: 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. 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. Simple geometric analysis yields that $\bar{u}$ and $\bar{u}_l$ are $\bar{d} = \sqrt{2\left(1 - \sqrt{(1-\gamma^2/4)}\right)}$ apart. The distance between $\bar{u}$ and $\bar{u}_r$ follows symmetrically. Having calculated the distance between $\bar{u}$ and $\bar{u}_l$, we are now in a position to characterize the probability of $\mathcal{E}_{no-optimal}$. But

$$\mathbb{P}(\mathcal{E}_{no-optimal}) = \mathbb{P}(V_* \notin \Xi) = \mathbb{P}\left(\|u_0 - \bar{u}\|_2 \geq \bar{d}\right) .$$

Note by Holder's inequality that, $\bar{d} \leq \|u_0 - \bar{u}\|_2 \leq \|u_0 - \bar{u}\|_1$, which implies that,

$$\mathbb{P}(\mathcal{E}_{no-optimal}) = \mathbb{P}\left(\|u_0 - \bar{u}\|_2 \geq \bar{d}\right) \leq \mathbb{P}\left(\|u_0 - \bar{u}\|_1 \geq \bar{d}\right) .$$

Note that $u_0 \sim \mathcal{N}(\bar{u}, \frac{\sigma^2}{d^2} I_d)$. Using Lemma 5 in Appendix D, we have that,

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

which results in a lower bound.

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, letting $C_2 = (d/\sigma)\,(\rho/(1-\gamma))$, and $\bar{u}_{max} = \max_{i=1,..,d} |\bar{u}_i|$, notice

$$\begin{aligned}
\mathbb{P}(u_0^\top a_i \geq \rho) &= \mathbb{P}(u_0^\top u_0 - u_0^\top u_0 + u_0^\top u_i \geq \rho) = \mathbb{P}(u_0^\top (u_0 - u_i) < u_0^\top u_0 - \rho) \\
&\geq \mathbb{P}\left(\|u_0\|_2 < \frac{u_0^\top u_0 - \rho}{\gamma} \mid u_0, u_i \in \Xi\right) = \mathbb{P}\left(\|u_0\|_2^2 - \gamma\|u_0\|_2 - \rho > 0 \mid u_0, u_i \in \Xi\right) \\
&= \mathbb{P}\left(\|u_0\|_2(1-\gamma) - \rho > 0 \mid u_0, u_i \in \Xi\right) = \mathbb{P}\left(\|u_0\|_2 > \rho/(1-\gamma) \mid u_0, u_i \in \Xi\right) \\
&\geq \mathbb{P}\left(|u_0^{max}| \geq \rho/(1-\gamma) \mid u_0, u_i \in \Xi\right) \geq \mathbb{P}\left((u_0^{max} - \bar{u}_{max}) \geq \rho/(1-\gamma) \mid u_0, u_i \in \Xi\right) \\
&\geq (\sqrt{4 + C_2^2} - C_2)\exp\left(-C_2^2/2\right)/2\sqrt{2\pi} ,
\end{aligned}$$

where the last inequality follows by the lower bound on tail probabilities of standard normal random variables (Duembgen 2010). This in-turn shows that with probability at least $(\sqrt{4 + C_2^2} - C_2)\exp\left(-C_2^2/2\right)/2\sqrt{2\pi}$, 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, by definition, that

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

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

$$\mathbb{E}_{U_0 \sim \mathcal{P}}\left[\mathcal{R}^{CB}(T, \rho, p, u_0)\right] = \mathbb{E}_{U_0 \sim \mathcal{P}}\left[\sum_{t=1}^{T} r_t(\rho, p, u_0)\right]$$

$$\begin{aligned}&\leq \mathbb{E}\left[\sum_{t=1}^{\top}(\mu(u_0^\top V_*)-\mu(u_0^\top a_t))+\mu(u_0^\top a_t)(1-\Pi_{t=1}^{\top}\mathbb{1}\{\Upsilon_t=0\})\right]\\&=\sum_{t=1}^{T}\mathbb{E}\left[(\mu(u_0^\top V_*)-\mu(u_0^\top a_t))\right]+\mathbb{E}\left[\mu(u_0^\top a_t)\left(1-\Pi_{t=1}^{\top}\mathbb{1}\{\Upsilon_t=0\}\right)\right].\end{aligned}$$

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

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

Now notice that our selection of the upper confidence around $\hat{u}$ depends on the link function μ. If μ is not the identity function than, our selection is the same as that of the GLM-UCB Algorithm of Filippi et al. (2010). Hence, following Theorem 2 in Filippi et al. (2010) , we have that

$$\mathcal{R}^{CB(\lambda,\gamma)}(T,\rho,p,u_0)\leq\tilde{C}d\log(sT)\sqrt{2T\log(2dT)}=\tilde{\mathcal{O}}\left(\sqrt{T}\right),$$

where $\tilde{C}:=(d+1)Y_{max}+\left(2\sqrt{3+2\log(1+2L^2/\lambda)}\kappa_\mu Y_{max}\right)/c_\mu$. Otherwise if μ is the linear, than our selection of the arm follows the OFUL algorithm of Abbasi-Yadkori et al. (2011) whose regret is given by

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

where we have used that $\|u_0\|_2\leq\bar{d}+1$ by step 1. This proves the final result. □

B.4. Dynamic updating of constrained set

Following Remark 2, we now analytically construct a problem instance to prove that dynamically updating the constrained product set can be outperformed by a fixed initial product selection. Let $U^{CB-R(\rho,p)}$ denote the re-optimizing policy with batch size $B=1$.

Proposition 3. *Consider $d=2$ with 3 products and let $u_0\sim\mathcal{N}(\bar{u},I_2)$. Also let $\dot{\mathcal{E}}=\{u_0:u_0\in U^{CB(\rho,p)}$ & $u_0\notin U^{CB-R(\rho,p)}\}$ denote the set of user realizations for which the static product set outperforms the dynamically updated product set. Then,*

$$\inf_{V_i,\rho,p,\bar{u}}\mathbb{P}_{u_0\sim\mathcal{P}}(\dot{\mathcal{E}})>0.$$

Proof of Proposition 3: We will prove this result by constructing a setting where the initial product selection outperforms dynamic updating of the constrained set. Let the prior mean on user latent features be $\bar{u}=[1,1/2]$ and note that the prior variance is given by I_2. Let product features be given by $V_1=[1,0],V_2=[0,1],V_3=[-1/\sqrt{2},1/\sqrt{2}]$. Let ρ and γ be selected so that at most two products remain in the constrained set and 1 of the available products is irrelevant for the customer. Finally, let $\mu(x)=x$ with error distributed as $\mathcal{N}(0,1)$. Then, clearly our static constrained set $\Xi=\{1,2\}$. Note that for for any customer realization with both components positive ($u_{01}>0$ & $u_{02}>0$), this product set is optimal; *i.e.,* product 3 is irrelevant and product 1 and 2 are relevant. Now, consider the utility of such a customer realization under the dynamic updating policy. Let $\hat{u}_t$ denote the posterior mean on customer features at time t. After round 1, the posterior mean is given by

$$\hat{u}_1 = \left[I_2 + [a_1^\top a_1]\right]^{-1} \left[\bar{u} + a_1^\top Y_1\right].$$

Given the prior on customer features, it is easy to observe that the first recommendation will be Product 1. Hence, $\hat{u}_t = [1/2 + Y_1/2, 1/2]$. Thus, the posterior mean only shifts in the first component while the second component remains the same. Furthermore, $Y_1 = u_{0_1} + \varepsilon$, where $u_{0_1} > 0$ is the real unknown user feature in dimension 1. Now consider the event $\mathcal{E} = \{\varepsilon < -(1 + u_{0_1})\}$. Under this event, the posterior mean becomes negative in dimension 1 and Product 1 becomes the least relevant product. Hence, dynamically updating the constrained set will lead to an exclusion of Product 1 (a relevant product) and an inclusion of Product 3 (an irrelevant product) from the updated set. Consequently, in round 2, there is a positive probability that the customer disengages from the platform if Product 3 is recommended. In contrast, with the fixed product set, the customer would remain engaged throughout the time horizon, since all recommendations would be restricted to the set of relevant products ($\{1, 2\}$). This proves the result. □

The construction in Proposition 1 shows that re-optimizing the constrained set does not always improve performance relative to a static product set; again, this is because noise in customer response can cause fluctuations to the product set that are harmful for engagement. Our numerical results in §5.1 complement this result in more general settings.

C. Practical Considerations

We now discuss a number practical considerations when using the Constrained Bandit. C.1 provides guidance on how to select the set diameter parameter γ to constrain exploration; C.2 shows that our algorithm is robust to modest misspecifications in the model parameters; C.3 shows that our results remain qualitatively similar when we only have distributional information on the model parameters (γ and ρ) across the customer population; C.4 shows that our algorithm still provides value when customers only temporarily disengage.

C.1. Selecting set diameter γ

We proved earlier 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 2). In this section, we explore this dependence in more detail and approximate the γ that maximizes the fraction of satisfied customers. First note that Step 2 in the analysis of Theorem 3 can be updated to show an explicit dependence on the prior on user latent features. In particular, letting $\bar{u}_{max} := \max_{i=1,..,d} \bar{u}_{max_i}$, we can show that when $1 > \gamma > 1 - \rho/\bar{u}_{max}$, then the fraction of customers who remain satisfied with the platform is lower bounded by

$$\left(1 - 2d \exp\left(-\left(1 - \sqrt{1 - \gamma^2/4}\right)/\sigma\right)\right) \left((\sqrt{4 + C_2^2} - C_2) \exp\left(-C_2^2/2\right)/2\sqrt{2\pi}\right),$$

where $C_2 = (d/\sigma)\left(\rho/(1-\gamma) - \bar{u}_{max}\right)$. Otherwise, when $\gamma \leq 1 - \rho/\bar{u}_{max}$, then the fraction of engaged satisfied customers is lower bounded by $.5\left(1 - 2d\exp\left(-\left(1 - \sqrt{1 - \gamma^2/4}\right)/\sigma\right)\right)$. Now notice that this is an increasing function of γ. Hence, when $\rho \leq \bar{u}_{max}$, we select

$$\gamma^* \approx 1 - \rho/\bar{u}_{max} > 0.$$

When $\rho > \bar{u}_{max}$, we choose the solution to

$$\gamma^* \approx \underset{0\leq\gamma\leq 1}{\arg\max} \left(1 - 2d\exp\left(-\left(1-\sqrt{1-\gamma^2/4}\right)/\sigma\right)\right)\left((\sqrt{4+C_1^2}-C_1)\exp\left(-C_1^2/2\right)/2\sqrt{2\pi}\right).$$

The problem above has no closed form solution but it is a single parameter optimization problem that can be solved using numerical optimization.

C.2. Robustness to misspecification of disengagement parameters

In Figure 5, we compare the performance of the CTS algorithm on the time of engagement when the tolerance threshold ρ is under-estimated by 5% (left), correctly estimated (center) and over-estimated by 5%. We find that the CTS algorithm is robust to misspecification in both cases, but over-estimation of ρ is preferable to under-estimation. Hence, when uncertain, we suggest selecting a larger value of the tolerance threshold.

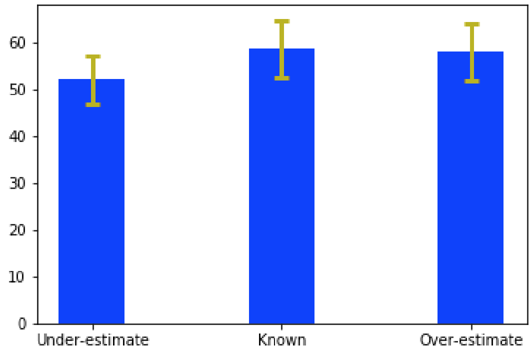

Figure 5 Performance of the CTS algorithm on the time of engagement metric when the tolerance threshold is under-estimated by 5% (left), correctly estimated (center) and over-estimated by 5%. We find that the CTS algorithm is relatively robust to overestimation but under-performs when ρ is under-estimated.

C.3. Prior distribution on customer tolerance and disengagement propensity

In practice, the disengagement parameters ρ and p may vary by customer; we now extend to the case where each customer's disengagement parameters are sampled from a known joint distribution $\dot{f}(\rho, p)$. Let f_p denote the marginal distribution of the disengagement propensity and f_ρ denote the marginal distribution of the relevance threshold. We assume that disengagement is salient for all customers: i.e $f_p(0) = 0$. Since Step 1 of Theorem 2 is only dependent on the size of the constrained set, the analysis remains the same as before. Step 2 of the analysis estimates a lower bound on the probability of engaged customers and naturally depends on the distribution of ρ. We re-evaluate this probability for this modified setting.

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. Noting that both have prior distributions, and letting $C_1(x) = (d/\sigma)\left(x/(1-\gamma)\right)$, we have that

$$\mathbb{P}_{u_0,\rho}(u_0^\top a_i \geq \rho) \geq \mathbb{P}_{u_0,\rho}\left(\|u_0\|_2(1-\gamma) - \rho > 0 \mid u_0, u_i \in \Xi\right) = \mathbb{P}_{u_0,\rho}\left(\|u_0\|_2 > \rho/(1-\gamma) \mid u_0, u_i \in \Xi\right)$$

$$
\begin{aligned}
&\geq \int_{x,y} \mathbb{P}_{u_0,\rho}\left(|u_0^{max}| \geq x/(1-\gamma) - \bar{u}_{max}\right) \dot{f}_{\rho,p}(x,y)dxdy \geq \int_x \mathbb{P}_{u_0,\rho}\left(|u_0^{max}| \geq x/(1-\gamma) - \bar{u}_{max}\right) f_\rho(x)dx \\
&\geq \int_x (\sqrt{4+C_1^2(x)} - C_1(x)) \exp\left(-C_1^2(x)/2\right)/2\sqrt{2\pi}dx \\
&= \mathbb{E}_{x\sim f_\rho}\left[\left(\sqrt{4+C_1^2(x)} - C_1(x)\right)\exp\left(-C_1^2(x)/2\right)/2\sqrt{2\pi}\right],
\end{aligned}
$$

where the last inequality follows by tail bounds on multivariate gaussians. Since, the rest of the proof of Theorem 2 continues to hold, we get that at least

$$
\left(1 - 2d\exp\left(-\left(1-\sqrt{(1-\gamma^2/4)}/\sigma\right)\right)\right)\mathbb{E}_{x\sim f_\rho}\left[\left(\sqrt{4+C_1^2(x)} - C_1(x)\right)\exp\left(-C_1^2(x)/2\right)/2\sqrt{2\pi}\right],
$$

fraction of customers remain satisfied on the platform. Hence, for any chosen γ, the distribution of the customer threshold plays an important role in estimating customer disengagement due to irrelevance.

We now examine this setting numerically. In Figure 6, we plot the fraction of engaged customers as a function of the constraint parameter when the underlying distribution on customer tolerance is uniform or truncated normal. We see the same tradeoff as before: very small sets do not contain the preferred products of any customers, but very large sets cause excessive exploration and disengagement. We also observe that varying ρ affects the fraction of engaged customers, but does not really affect the optimal set diameter γ^*. Similar to Appendix C.1, we can use our earlier theoretical analysis to estimate a good choice of γ that maximizes the fraction of satisfied customers under a given prior.

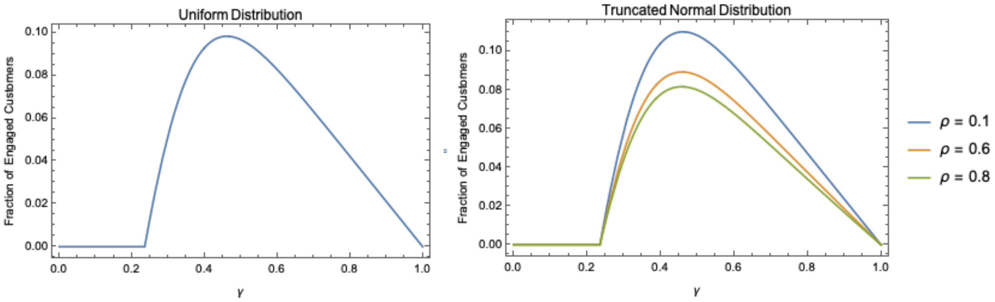

Figure 6 **Minimum fraction of engaged customers as a function of the constraint threshold parameter for uniform (left) and truncated normal (right) distributions.**

C.4. Temporary disengagement

We now consider the model with temporary disengagement described in §3.4. Unlike Proposition 1, when disengagement is sufficiently temporary, one *can* obtain sublinear regret over all customers.

PROPOSITION 4. *Under temporary disengagement with $\delta \leq 1$, the regret of any non-anticipating policy is*

$$
\mathcal{R}(T,\rho,\delta,u_o) \geq CT^{\max\left\{\frac{1}{2},\delta\right\}},
$$

where C is a constant independent of T.

We omit the proof since it follows directly from combining the standard lower bound in bandit problems (without disengagement) and our construction in the proof of Proposition 1. In particular, when $\delta = 0$ (customers disengage for a negligible time period), the classic lower bound applies and yields at least $\sqrt{T}$ regret. For nonzero δ, since the customer may disengage on the first time step with positive probability, the lower bound on regret due to disengagement is at least T^{δ}.

It is also easy to observe that when disengagement is relatively short (*i.e.,* $\delta \leq 1/2$), consistent bandit algorithms achieve the regret rates as in the classical setting without disengagement. Yet, this is an asymptotic claim; we now show numerically that there is still value in constraining exploration even when $\delta \leq 1/2$.

In particular, we compare Thompson Sampling (TS), greedy Bayesian updating (MLE), and our approach of constrained Thompson Sampling (CTS). To model the information from disengagement, we re-solve the IP with an exclusion constraint that ensures that the product that caused disengagement is not recommended again to the same customer; a similar exclusion constraint is also added for the MLE and the TS algorithm for a fair comparison.

The top row of Figure 7 plots engagement and cumulative regret when $\delta = 0.5$, *i.e.,* when the theory prescribes that disengagement is asymptotically negligible. We see that the CTS algorithm still considerably outperforms the benchmark algorithms on both metrics. The bottom row plots the same measures when disengagement is even more temporary: $\delta = 0.25$. We see that even in this setting, CTS continues to outperform the benchmark algorithms, albeit by a narrower margin.

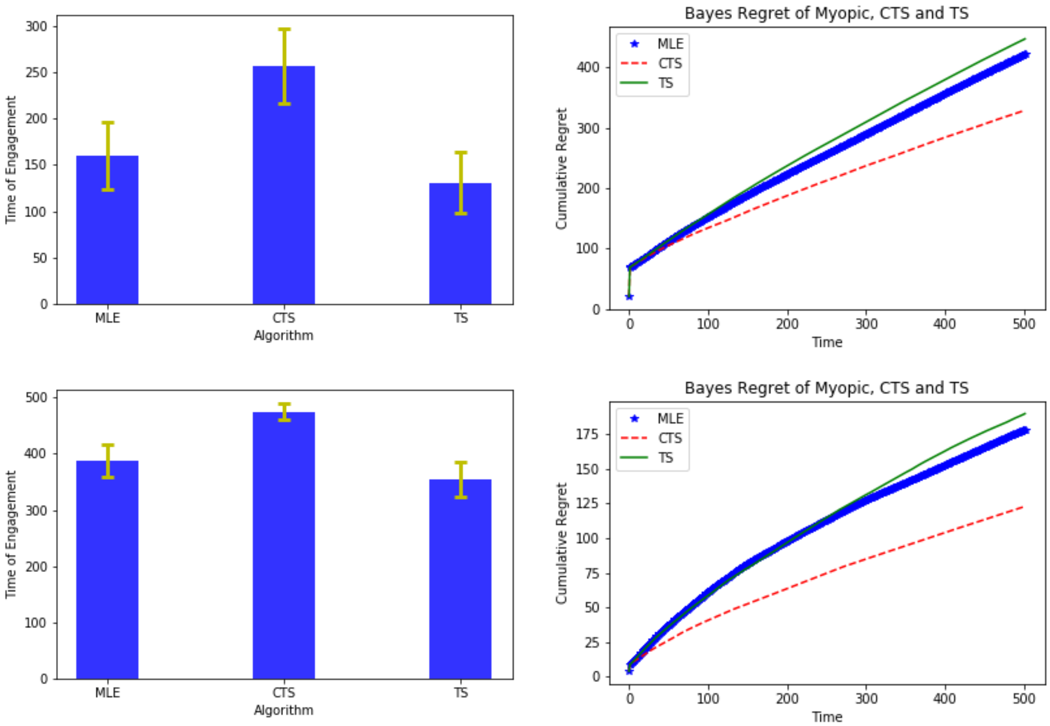

Figure 7 Total time of disengagement and Bayes regret averaged over 100 customers when customers leave for $\sqrt{T}$ periods (top) and when they leave for $T^{1/4}$ periods (bottom).

D. Auxiliary Results

Greedy Bayesian Updating: Algorithm 3 formally states the greedy Bayesian updating algorithm from §3.3. Note that when μ is the identity function (linear case), step 2 of the algorithm can be simplified to set

$$\hat{u}_{t+1} = \left(a_{1:t}^\top a_{1:t} + \frac{\xi^2}{\sigma^2} I \right)^{-1} \left(a_{1:t}^\top Y_{1:t} \right) .$$

Algorithm 3 Greedy Bayesian Updating (GBU)

```
Initialize and recommend a randomly selected product.
for $t \in [T]$ do
    Observe customer utility, $Y_t = \mu(u_0^\top a_t) + \varepsilon_t$.
    Update customer feature estimate, $\hat{u}_{t+1} = \sum_{k=1}^{t} (Y_k - \mu(a_k^\top \hat{u}_t)) a_k = 0$.
    Recommend product $a_{t+1} = \arg\max_{i=1\dots,n} \hat{u}_{t+1}^\top V_i$.
end for
```

Technical results: We now state a useful lemma.

Lemma 5. *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: The proof follows from simple application of the pigeon-hole principle and tail bounds on multivariate normal variables. □

Definition 4 formally defines consistent bandit algorithms.

Definition 4 (Lattimore and Szepesvari 2016). *A policy π belongs in the class of consistent bandit algorithms Π^C 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*:

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