Algorithmic Content Selection and the Impact of User Disengagement

Emilio Calvano[1], Nika Haghtalab[2], Ellen Vitercik[3], and Eric Zhao[4]

[1]University of Rome, emilio.calvano@gmail.com
[2]University of California, Berkeley, nika@berkeley.edu
[3]Stanford University, vitercik@stanford.edu
[4]University of California, Berkeley, eric.zh@berkeley.edu

October 18, 2024

Abstract

The content selection problem of digital services is often modeled as a decision-process where a service chooses, over multiple rounds, an arm to pull from a set of arms that each return a certain reward. This classical model does not account for the possibility that users disengage when dissatisfied and thus fails to capture an important trade-off between choosing content that promotes future engagement versus immediate reward. In this work, we introduce a model for the content selection problem where dissatisfied users may disengage and where the content that maximizes immediate reward does not necessarily maximize the odds of future user engagement. We show that when the relationship between each arm's expected reward and effect on user satisfaction are linearly related, an optimal content selection policy can be computed efficiently with dynamic programming under natural assumptions about the complexity of the users' engagement patterns. Moreover, we show that in an online learning setting where users with unknown engagement patterns arrive, there is a variant of Hedge that attains a $\frac{1}{2}$-competitive ratio regret bound. We also use our model to identify key primitives that determine how digital services should weigh engagement against revenue. For example, when it is more difficult for users to rejoin a service they are disengaged from, digital services naturally see a reduced payoff but user engagement may—counterintuitively—increase.

1 Introduction

Content selection is an important problem for online platforms that encompasses, for example, the use of recommendation algorithms by social media apps. In the content selection problem, a system (the app) repeatedly selects pieces of content to provide to a user. The content is chosen from a pool of options with the goal of maximizing some notion of long-term payoff, such as the cumulative ad revenue that is realized from a user. Decision processes like multi-armed bandits (MAB) are commonly used to model these problems, where each piece of content is abstractly represented as an arm, and the amount of revenue earned by serving the content is represented as the arm's reward. Such classical models of content selection implicitly rely on the simplifying assumption that user engagement is a given: given a captive user, the decision maker's only concern is to maximize revenue. In practice, user engagement is a variable quantity that depends intimately on the user's

previous experiences with the app and the user's own patterns of engagement; a user may open an app less frequently if previous interactions have built up a sense of dissatisfaction.[1]

This user engagement dimension of content selection—which is often overlooked in algorithmic models—involves a complex tradeoff between multiple objectives, such as revenue, engagement, and user satisfaction. Existing methods for multi-objective learning are ill-suited for addressing these tradeoffs. The foremost challenge is that building user engagement is a stateful, noisy, and long-time-horizon problem—the type of adaptive decision process that one might ordinarily approach with reinforcement learning. This is in contrast to the problem of maximizing revenue per interaction, which does not typically require long-term planning. Another challenge is that user engagement is a stochastic process where the results of future interactions are heavily confounded with previous interactions: if a user is dissatisfied with an app, the app has fewer opportunities to influence the user's opinion about the app. These are core challenges in the content selection problem which we address directly in the algorithm design problem.

Prior research has taken a first step towards addressing variable engagement in user interactions [3, 5, 18]. However, to fully encapsulate the challenges associated with content selection, our work deviates from two common simplifying assumptions made by existing work. First, we consider content landscapes for which the generated revenue need not fully align with user satisfaction. Thus, our optimal policies must account for rich tradeoffs between multiple objectives, such as user engagement and revenue maximization. On the other hand, when engagement is directly aligned with revenue, as in previous work, the optimal policy need not make such tradeoffs. Secondly, our model allows users to disengage by temporarily leaving the platform, during which time the platform cannot impact user satisfaction. Therefore, we must account for the statistical correlations that arise between a user's previous choices to disengage and the likelihood that they will disengage in the future. On the other hand, in previous work, either users disengage permanently, or the platform retains the ability to influence the disengaged user's state. Either of these assumptions obviates the need to account for correlations across user states.

In this work, we introduce a general model for content selection that captures the engagement-revenue trade-off and prove efficient algorithms for computing optimal policies and obtaining online learning guarantees for adversarial sequences of unknown users. The model captures the trade-off between investing in the quality of the user experience and generating revenue by allowing for situations where improving user experience is inversely correlated with revenue gains. The model is also expressive enough to reflect that an app lacks influence on a user's opinion when said user is disengaged, that users may return after disengaging, and that friction can make it relatively more difficult for users to re-engage when disengaged. The main contributions of this paper are as follows:

Efficient content selection algorithms for optimal policy computation and online learning. In each timestep of our model, a user chooses whether to interact with an app with a probability that is given by some function—which we refer to as the user's engagement pattern—of the user's cumulative satisfaction with the app. We show that for a user whose engagement pattern features only k possible levels (probabilities) of engagement, an exact optimal policy can be computed using dynamic programming with a runtime of $O(k^2)$ (Theorem 3.3). We also study an online learning setting where a series of T—potentially adversarially selected—users with unknown engagement patterns are introduced to an app, which chooses a content selection policy for each user that arrives and is provided with a realization of the policy's cumulative revenue. We prove that in this online learning setting, in regimes with limited content, there is an algorithm that can achieve a sublinear

[1]The meaning of 'reduced engagement' depends on the application. It could be defined as uninstalling the app, becoming inactive, or merely closing an active session. These are equivalent for our purposes.

regret of $O(\sqrt{T})$ (Theorem 3.6). Meanwhile, in regimes where initial user engagement is high, there is an algorithm that can achieve an approximate regret ratio of $\frac{1}{2}$ (Theorem 3.9). These results rely only on a mild linearity condition on the content landscape: that each piece of content's expected revenue and effect on user satisfaction follow a linear relationship, which can be negative.

The key idea that allows us to obtain these efficient algorithms is the observation that a content selection problem with variable user engagement can be reduced to content selection with a captive audience but where the discount rate varies depending on the user's state (Theorem 2.2).

Modified demand elasticity. We study an analogue of the classical notion of demand elasticity [6, 11] in our model, which we refer to as modified demand elasticity. Just as demand functions are classically used to abstract away other market forces such as consumers and competition, the modified demand function exogenizes the effect of outside factors on user engagement. We show that the shape of the modified demand function fully explains the impact of model primitives (namely the discount factor, content landscape, and features of the engagement pattern) on how apps invest in user experience.

Using modified demand elasticity to analyze friction and alignment. We also use modified elasticity to study how friction affects user engagement and to understand when app creators are incentivized to invest more in user satisfaction. Friction is a primitive in our model that describes the relatively lower probability of a user engaging with an app after being disengaged versus when previously engaged. It is easy to verify that higher friction strictly reduces the payoff that an app can expect to obtain. However, we show that if apps use optimal content selection policies, the amount of user-app engagement can counter-intuitively increase with friction (Example 4.1). This phenomenon is due to friction increasing (what the app perceives to be) the user's modified demand elasticity, which increases the app's incentive to invest in user satisfaction. We also prove that, when modified demand elasticity is high, app creators are incentivized to invest more in building user satisfaction such that user engagement remains high (Theorem C.3).

1.1 Related work

Recent works in bandit learning have also aimed to model decision-making in settings that require both optimizing for immediate rewards and for future engagement opportunities. Ben-Porat et al. [3], Cao et al. [5], and Yang et al. [18] propose bandit learning models where every arm pull yields not only a reward but also causes the episode to end with a certain probability. These models are largely motivated by the design of engagement-aware algorithms for recommendation systems. These models differ from our work in two fundamental ways. First, that user disengagements are permanent events and, more critically, each arm's revenue is directly aligned with each arm's effect on user engagement. That is, the algorithm's revenue is either directly defined as being the number of user engagements or the probability an arm causes disengagement is decreasing in the reward of the arm. In contrast, our model captures the potential for user re-engagement and allows conflict between the objectives of revenue and engagement maximization. Other works have studied bandits with more general notions of state that can be affected by prior arm pulls [e.g., 8, 9, 10], but where the user's state affects arm payoffs rather than engagement or horizon length.

Beyond bandit learning, Zhang et al. [19] studies a decision-making problem where the sole objective is revenue maximization, but subject to a hard constraint that users have a positive expected future utility. This constraint-based approach to balancing engagement cannot capture partial disengagement risk and, by basing engagement off expected future utility rather than a user's previous experiences, both assumes that users are omniscient about the app's intentions and

means that an app's previous actions has no baring on it's future opportunities for engagement. Pacchiano et al. [15] study a similar constraint-based setting in a more generic online learning setting with linear constraints. In contrast to the models of Ben-Porat et al. [3], Cao et al. [5], and Yang et al. [18], the models of Zhang et al. [19] and Pacchiano et al. [15] allow for misalignment between satisfying engagement constraints and reward maximization.

2 Modeling Content Selection under Engagement-Revenue Tradeoffs

Basic model. We now present a basic version of our content selection model, which builds on the classical model where an app chooses, at every timestep, a piece of content from a set of options that each provide a certain amount of revenue.

1. A user interacts with an app at the first timestep, $t = 1$.
2. If the user chooses to interact with the app at timestep t, the app picks a piece of content i_t from a set $\mathcal{I}$ to show the user. Each content $i \in \mathcal{I}$ earns the app a revenue described by the random variable R_i and provides the user an experience represented by the random variable E_i, where both E_i and R_i are supported on $\mathbb{R}$. We denote realized revenue and user experience by $r_t \sim R_{i_t}$ and $e_t \sim E_{i_t}$ respectively.
3. If the user does not interact with the app at timestep t, the app has no actions available, earns no revenue, and makes no impression on the user, i.e., $i_t = \emptyset$, $r_t = 0$, and $e_t = 0$. We indicate the event that a user interacts with the app at timestep t with the variable $s_t = \mathbb{1}[\text{User Interacts}]$.
4. The user determines whether they want to engage with the app in the next timestep based on a summary $x_{t+1} = \phi(e_1, \ldots, e_t)$ of its previous app experiences; we will fix $\phi(e_1, \ldots, e_t) = \sum_{\tau=1}^{t} e_\tau$ for exposition but note that our results will also hold for other choices of ϕ (see the paragraph titled *Generalizations*). If the user is already on the app, i.e. $s_t = 1$, they continue usage with a probability that is a monotonically non-decreasing function f of x_t; that is, $s_t \sim \text{Bernoulli}(f(x_t))$. If the user is not already on the app, this engagement probability is scaled down by a constant $1 - c \in [0, 1]$; that is, $s_t \sim \text{Bernoulli}((1 - c) \cdot f(x_t))$.

We refer to f as the user's *demand function* and the constant c as the *friction parameter*, where larger c corresponds to greater friction. We refer to the summary of a user's previous interactions x_t as the *user's state*; since user demand $f(x_t)$ is increasing in user state x_t, one can understand the user's state to roughly reflect a user's satisfaction with an app or tendency to use an app. For technical reasons, we assume the sets $\{\mathbb{E}[E_i]\}_{i \in \mathcal{I}}$ and $\{\mathbb{E}[R_i]\}_{i \in \mathcal{I}}$ are compact.

App behaviors in the model. A content selection policy π maps from a transcript of prior interactions, $H_T = \{(s_t, r_t, e_t, i_t)\}_{t \in [T]}$, to a distribution over content; that is, $i_{T+1} \sim \pi(H_T)$. The objective of the app's content selection policy is to maximize long-term payoff

$$J(\pi) := \mathop{\mathbb{E}}_{\{(s_t, r_t, e_t, i_t)\}_t} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right], \tag{1}$$

a time-discounted sum of app revenues where $\gamma \in (0, 1)$ is the app creator's *discount factor* and describes how patient/forward-sighted the app should be. We note that in Equation 1, the expectation is taken over the transcript, which is a stochastic process.

Policies that are an arbitrary function of history are unwieldy. Fortunately, it suffices for apps to only consider *simple policies*: policies that choose content at time t using only the user's state x_{t-1}. In more detail, a policy π is simple if there exists $g : \mathbb{R} \to \mathcal{I}$ where for every transcript $H = \{(s_t, r_t, e_t, i_t)\}_{t\in[T]}$, we have $\pi(H) = g(\sum_{\tau=1}^{t} e_\tau)$. This is formalized in the lemma below.

Lemma 2.1. *If there is an optimal policy for the app creator, there is also a simple optimal policy.*

We use the indexing $w^{(t)}$ to denote the value of a variable w at the t-th user-app interaction. Observe that one can write the content $i^{(t)}$ chosen by any simple policy as a function of $x^{(t-1)}$.

Interpretation in the context of classical models. This model arises naturally from two main abstractions of user-platform interactions:

(a) The platform's revenue accumulates additively as a result of arm pulls, where each arm corresponds to a stationary distribution.

(b) The stateful variable that determines the probability of user engagement (e.g., user satisfaction) accumulates additively as a result of arm pulls, where each arm corresponds to a stationary distribution.

Abstraction (a) is shared by many recommendation system models and is traditionally motivated by thinking about arm rewards as literal money. Our model augments the classical recommendation system model by applying a similar abstraction for user engagement. This is abstraction (b), which treats user satisfaction like revenue in that it accumulates from the sum of previous arm pulls. Abstraction (b) is motivated by prior literature in operations research [e.g., Baucells and Sarin [2]] and recommendation systems [e.g., Leqi et al. [10]] that similarly model user state (e.g., satisfaction, satiation) as evolving linearly. More generally, our modeling of user-platform engagement as a decision made by the user is inspired by dynamic mechanism design literature that model platforms as aiming to create an environment where users are incentivized to participate/engage with the platform. Related works studying recommendation systems under participation constraints [19] and from the perspective of addiction [7] have similarly built on this perspective. More broadly, accounting for the fact that users re-engage is core to the empirically-driven design of human-computer interfaces and assessment of user engagement [14].

The demand function f in our model is also closely connected to existing notions of demand in markets with prices. Traditionally, it is common to abstractly represent a consumer's decision-making as to how much of a good to purchase as an exogenous function mapping from the firm's action (price set) to the consumer's action (amount to purchase). This model similarly represents the user's decision-making as a demand function mapping user satisfaction to engagement probability. This means that user satisfaction—which is controlled by the platform—is analogous to the price set by a firm; and the user's engagement probability is analogous to the quantity of goods purchased. This analogy becomes literal if we consider a setting where the platform is a grocery store choosing what prices to set and the user is a shopper who chooses whether to visit the store based on the prices it experienced on previous visits—a situation fully captured by our model. One can also think of the platform itself as an experience good, for which the demand function describes the user's estimation of its value.

Generalizations of the basic model. Our basic model can be generalized in several ways. First, our model can be generalized to define user satisfaction as a discounted sum of previous arms pulls $x^{(t+1)} = \sum_{\tau=1}^{t} \gamma^\tau e^{(\tau)}$ or as the mean of previous arm pulls $x^{(t+1)} = \frac{1}{t}\sum_{\tau=1}^{t} e^{(\tau)}$. For example,

consider when a user decides whether to visit a store based on the average price they have experienced at the business; this example is fully captured by our model by defining user satisfaction as the mean of previous experienced prices. Our model can also be generalized to define user experiences, i.e. the supports of $\{E_i\}_{i\in\mathcal{I}}$, as vectors in $\mathbb{R}^d$ and the user's demand function $f : \mathbb{R}^d \to [0, 1]$ to be monotonically non-decreasing in the product order of $\mathbb{R}^d$. For clarity, we present the results in this paper for our basic model. However, all of the results we present also directly extend to these generalizations.

On friction. We model friction as a multiplicative factor because friction is generally understood as proportionally decreasing engagement, e.g. where engagement is reduced by 50%, rather than an absolute decrease. However, our results generalize readily to other notions of friction—including additive forms of friction or where re-engagement probability is allowed to depend on how long it's been since the user was last on the platform. More concretely, our model of friction can be generalized such that such, instead of scaling down user demand, friction can arbitrarily affect interaction probabilities; that is, the model defines a set of demand functions $\{f_n\}_{n\in\mathbb{Z}}$ where the user interacts with probability $f_n(x_t)$ if the last interaction was n timesteps prior. All technical results we present directly extend to this generalization.

For concrete examples of factors that may cause the friction to differ between two platforms, consider when:

1. One digital service is a native iPhone application able to send push notifications, whereas the other digital service is a web application that cannot send push notifications. The latter may experience greater friction due to having fewer options to communicate with disengaged users.

2. Two digital services, called A and B, are competing for the same pool of users. Let digital service A be more addictive than the other. Then service B will experience greater friction, as when it loses a user to service A, the user is less likely to return due to A's addictiveness. We explore this situation in more detail in Example C.11.

3. One digital service has more inelastic demand than a digital service. For example, one might expect that a social network used for work and not out of personal desire—meaning that engagements are primarily driven by external motivation—may be subject to less friction than a social network used for personal pleasure.

2.1 From Variable Engagement to Variable Discount Rates

In order to tractably approach the computation of optimal policies in our model, we make the observation that, from an app's perspective, content selection with variable user disengagement is equivalent to content selection without user disengagement but with variable discount rates. This reduction follows from the observation that a user not interacting with an app for k timesteps is equivalent to the app creator's discount rate decreasing from γ to γ^{k+1}. Thus, when viewing the app creator's objective through this equivalent lens, we will allow the discount factor to vary according to a function $\widetilde{f}$ of the user's state. Since the number of timesteps that pass between user-app interactions follows the geometric distribution, the following theorem uses the geometric distribution's moment-generating function to define these variable discount rates.

Theorem 2.2. *For any simple policy π, the app creator's objective value can be written as*

$$J(\pi) = \mathop{\mathbb{E}}_{\left\{(s^{(t)},r^{(t)},e^{(t)},i^{(t)})\right\}_t}\left[\sum_{t=1}^{\infty} r^{(t)} \prod_{\tau=2}^{t} \gamma\widetilde{f}(x^{(\tau)})\right],$$

which uses a variable discount rate in which

$$\widetilde{f}(x) \coloneqq f(x) + \frac{1 - f(x)}{1 - \gamma(1 - (1 - c)f(x))}(1 - c)f(x)\gamma.$$

Proof. First, we define a function that maps from each timestep to the number of interactions that have occurred up to and including that timestep: $\mathsf{NumInteractions}(t) = \sum_{\tau=1}^{t} s_\tau$. We also use $\mathbb{T} = \{t \in \mathbb{Z}_+ \mid s_t = 1\}$ to denote the timesteps where user-app interactions took place. Observe that $\mathsf{NumInteractions}$ is bijective on the domain $\mathbb{T}$. We will now rewrite Equation 1 as a series supported only on $\mathbb{T}$. Since there is no revenue, i.e. $r_t = 0$, on timesteps where the user did not engage the app, i.e. $t \notin \mathbb{T}$, we can write Equation 1 as

$$J(\pi) = \mathop{\mathbb{E}}_{\{s_t, r_t, e_t, i_t\}_t(\pi)} \left[\sum_{t=1}^{\infty} s_t \gamma^{t-1} r_t \right] = \mathop{\mathbb{E}}_{\{s_t, r_t, e_t, i_t\}_t(\pi)} \left[\sum_{t \in \mathbb{T}} \gamma^{t-1} r_t \right].$$

Since we can invert the mapping $\mathsf{NumInteractions}$ on the domain $\mathbb{T}$, the inverse $\mathsf{NumInteractions}^{-1}$ is well-defined in the above series: $\left\{\mathsf{NumInteractions}^{-1}(t) \mid t \in \mathbb{T}\right\} = \mathbb{Z}_+$. We can therefore re-index

$$J(\pi) = \mathop{\mathbb{E}}_{\{s_t, r_t, e_t, i_t\}_t(\pi)} \left[\sum_{t \in \mathbb{T}} \gamma^{t-1} r_t \right] = \mathop{\mathbb{E}}_{\{s_t, r_t, e_t, i_t\}_t(\pi)} \left[\sum_{t=1}^{\infty} \gamma^{\mathsf{NumInteractions}^{-1}(t) - 1} r^{(t)} \right].$$

Recall that the indexing $r^{(t)}$ refers to the realized revenue at the t-th interaction, rather than the t-th timestep. By linearity, we can use telescoping to simplify

$$J(\pi) = \sum_{t=1}^{\infty} \mathop{\mathbb{E}}_{s_t, r_t, e_t, i_t} \left[r^{(t)} \gamma^{\mathsf{NumInteractions}^{-1}(t) - 1} \right] = \sum_{t=1}^{\infty} \mathop{\mathbb{E}}_{s_t, r_t, e_t, i_t} \left[r^{(t)} \prod_{\tau=2}^{t} \gamma^{w^{(\tau)}} \right],$$

where $w^{(\tau)} = \mathsf{NumInteractions}^{-1}(\tau) - \mathsf{NumInteractions}^{-1}(\tau - 1)$. Here, we have used the fact that the first timestep always corresponds to a user-app interaction, so $\mathsf{NumInteractions}^{-1}(1) = 1$.

The random variable $w^{(\tau)}$ can be understood as the amount of time that passes between the user's $(\tau - 1)$th and τth interactions with the app. Since the app is following a simple policy, $w^{(\tau)}$ is independent of $w^{(\tau - 1)}$ conditioned on the user's state after the $(\tau - 1)$th interaction, i.e. $x^{(\tau)}$.

If there is no friction in our model, $w^{(\tau)}$ is distributed according to the geometric distribution with parameter $f(x^{(\tau)})$. A geometric distribution with parameter p describes the number of coins with heads probability p that need to be flipped before a heads is flipped. When there is friction c, $w^{(\tau)}$ is distributed according to the non-homogeneous geometric distribution with the probability mass function $\Pr(k) = f_k(x^{(\tau)}) \prod_{\tau=1}^{k-1} f_\tau(x^{(\tau)})$ defined on the support $k \in \mathbb{N}$, where $f_1 = f$ and $f_n = cf$ for all $n > 1$. We will write this distribution with the shorthand $\mathrm{Geo}(p, c)$, One can compute the moment generating function of this non-homogeneous geometric distribution to be

$$\mathop{\mathbb{E}}_{X \sim \mathrm{Geo}(p,c)} \left[\exp(tX)\right] = p \cdot \exp(t) + \frac{1 - p}{1 - \exp(t)(1 - (1 - c)p)}(1 - c) \cdot p \cdot \exp(t)^2$$

for $t < -\ln(1 - (1 - c)p)$, which provides a simplified way in which to write the expectation

$$\mathbb{E}\left[\gamma^{w_t} \mid x_t\right] = f(x_t)\gamma + \frac{1 - f(x_t)}{1 - \gamma(1 - (1 - c)f(x_t))}(1 - c)f(x_t)\gamma^2.$$

Observe that $\mathbb{E}\left[\gamma^{w_t} \mid x_t\right] = \gamma \cdot \widetilde{f}(x_t)$. By the law of total expectation, we therefore recover

$$J(\pi) = \sum_{t=1}^{\infty} \mathop{\mathbb{E}}_{s_t, r_t, e_t, i_t} \left[r^{(t)} \prod_{\tau=2}^{t} \mathbb{E}\left[\gamma^{w^{(\tau)}} \,\middle|\, x^{(\tau)} \right] \right] = \mathop{\mathbb{E}}_{\{s_t, r_t, e_t, i_t\}_t} \left[\sum_{t=1}^{\infty} r^{(t)} \prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right].$$

□

3 Optimal Content Selection in Linear Settings

Solving the content selection problem in our model involves solving a long-term optimization problem. Every user interaction requires weighing trade-offs between harvesting immediate revenue and investing in future engagement. Moreover, user engagement is an adaptive stochastic process that is highly correlated with earlier content choices and user-app interactions. These factors significantly complicate optimizing and learning content selection policies. Surprisingly, if an app's content landscape demonstrates a linear relationship between revenue and user experiences, it is still possible to provide a simple description of the app's optimal content policy.

Linear setting. We say that an app creator's content decision problem is linear if the following holds. The content available to an app is represented by an interval $\mathcal{I} = [-K, K]$ for some $K > 0$. Each content $i \in \mathcal{I}$ provides a deterministic user effect $E_i = C_E - i$ and deterministic revenue $R_i = C_R + i$ where $C_E \in [0, K)$ and $C_R \geq 0$ are constants representing drift.

We note that this assumption does not imply that the model's overall dynamics are linear. First, the user demand function can be arbitrarily complex, which means—for example—that one can still model non-linear diminishing returns to increasing user satisfaction by choosing a sigmoidal demand function f. Second, the linearity assumption captures important classical settings. For example, an important case of the linear setting is where the platform is setting direct prices for goods, and a user decides whether to visit the platform based on its historical prices.

We also say a user has a demand function of complexity k if it is a piecewise constant function with k pieces; that is, there exists $b_1, \ldots, b_k \in \mathbb{R}$ such that f is constant on intervals $(-\infty, b_1), [b_1, b_2), \ldots, [b_{k-1}, b_k), [b_k, \infty)$. That is, the user's engagement pattern has k discrete levels, each corresponding to a given probability of engagement.

In linear settings, we can show that under an optimal policy, the frequency of app-user interactions stabilizes relatively fast. Moreover, an app's investment in user engagement—and by extension its behavior—is largely characterized by where user demand stabilizes, i.e., $f(x_\infty)$.

Lemma 3.1. *In the linear setting where the user demand function has a complexity of $k < \infty$, there is an optimal simple policy for the app creator that satisfies all of the following characteristics:*

- *The sequence $x_1, x_2, \ldots$ of user states is monotonic.*
- *The limit $x_\infty = \lim_{t\to\infty} x_t$ exists and is either at a discontinuity of f or negative infinity.*
- *(When $x_\infty = -\infty$) The app always shows the highest-revenue content, i.e. $i = K$.*
- *(When x_∞ is a discontinuity of f) The user state x_∞ will be reached within $k + n + 1$ interactions (i.e., $x_{n+k+1} = x_\infty$) where n is the smallest number of interactions in which x_∞ can be reached in some policy (i.e., there exists a policy where $x_n = x_\infty$). Moreover, if $n = 1$, $x_2 = x_3 = \cdots = x_\infty$.*

An important step towards proving Lemma 3.1 is arguing that the trajectory of optimal policies demonstrate structure with respect to discontinuities in one's demand function.

Lemma 3.2. *Consider an optimal simple policy π. For any subsequence of the policy's user state trajectory, $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$, where $x^{(2)}, \ldots, x^{(T)}$ are all not discontinuities of f:*

1. *If the policy π is not maximally increasing user state at the last step of the subsequence, i.e. $\pi(x^{(T)}) > -K$, then all previous actions in the subsequence should be maximally decreasing user state $\pi(x^{(1)}) = \pi(x^{(T-1)}) = K$.*
2. *If the policy π is not maximally decreasing user state at the first step, that is $\pi(x^{(1)}) < K$, then all later actions in the subsequence should be maximally increasing user state $\pi(x^{(2)}) = \pi(x^{(T)}) = -K$.*

3.1 Computing Optimal Content Selection Policies

An optimal policy can be efficiently computed in linear settings with dynamic programming. The following result builds on our Lemma 3.1's characterization of optimal policies in linear settings

Theorem 3.3. *In the linear setting where the user demand function has a complexity of $k < \infty$, an optimal app policy can be computed in a runtime of $O(k^2)$.*

Proof. To witness this claim, we construct Algorithm 1, a dynamic programming algorithm that estimates the optimal strategies for moving user engagement between every two pair of discontinuities in the user demand function.

Algorithm 1 Dynamic programming algorithm for computing app policies (Theorem 3.3).

Initialize: Dictionaries Trajs^*, V^*;
for discontinuity $d \in D \cup \{0\}$ where $x \leq 0$, in ascending order **do**
  Let $V_d^* = \widehat{v} + \frac{\gamma_T(C_R+K)}{1-\widehat{\gamma}\widetilde{f}(\min D)}$, $\text{Trajs}_d^* = \text{Traj} + [K]^*$ with $\widehat{v}, \widehat{\gamma}, \text{Traj} \leftarrow \text{GetPayoff}(d, \min D)$;
  for $d' \in D$ where $d' < d$, in ascending order **do**
    Let $v' = \widehat{v} + \widehat{\gamma} \cdot V_{d'}^*$ with $\widehat{v}, \widehat{\gamma}, \text{Traj} \leftarrow \text{GetPayoff}(d, d')$;
    If $v' > V_d^*$, let $V_d^* = v'$ and $\text{Trajs}_d^* = \text{Traj} + \text{Trajs}_{d'}^*$;
  end for
  If $\frac{C_R+C_E}{1-\gamma\widetilde{f}(d)} > V_d^*$, let $V_d^* = \frac{C_R+C_E}{1-\gamma\widetilde{f}(d)}$ and $\text{Trajs}_d^* = [C_E]^*$;
end for
for discontinuity $d \in D \cup \{0\}$ where $x \geq 0$, in descending order **do**
  If $d \neq 0$, let $V_d^* = \frac{C_R+C_E}{1-\gamma\widetilde{f}(d)}$ and $\text{Trajs}_d^* = [C_E]^*$;
  for $d' \in D$ where $d' > d$, in descending order **do**
    Let $v' = \widehat{v} + \widehat{\gamma} \cdot V_{d'}^*$ with $\widehat{v}, \widehat{\gamma}, \text{Traj} \leftarrow \text{GetPayoff}(d, d')$;
    If $v' > V_d^*$, let $V_d^* = v'$ and $\text{Trajs}_d^* = \text{Traj} + \text{Trajs}_{d'}^*$;
  end for
end for
Return Trajs_0^*;

Algorithm 1 invokes as a subroutine the following procedure for calculating trajectory payoffs.

Algorithm 2 Algorithm $\text{GetPayoff}(x, x^*, f)$.

Input: starting point x, ending point x^*, and demand function f;
Initialize: payoff $v_0 = 0$, discount $\gamma_0 = 1$, and position $x_0 = x$;
for $t \in [T]$ where $T = \left\lceil \frac{|x^*-x|}{K+C_E} \right\rceil$ is the timesteps to reach x^* from x **do**
  if $x^* < x$ **then**
    Take action $i_t = \min\{K, x_{t-1} - x^* + C_E\}$ to history Traj;
  else if $x^* > x$ **then**
    Take action $i_t = \begin{cases} \delta & \delta > 0 \wedge t = 1 \\ -K & \text{otherwise} \end{cases}$ where $\delta = x^* - x \bmod K + C_E$;
  end if
  Update $v_t = v_{t-1} + \gamma_{t-1} \cdot (C_R + \eta)$, $x_t = x_{t-1} + C_E - \eta$, and $\gamma_t \leftarrow \gamma_{t-1} \cdot \gamma\widetilde{f}(x_t)$;
end for
return Payoff v_T, discount γ_T, and trajectory $i_1, \ldots, i_T$;

We first prove some intermediate technical facts about Algorithm 1.

Fact 3.4. *For every discontinuity $d \leq 0$ of f, if $\pi^*(d) \geq C_E$, then Trajs_d^* describes the trajectory of the optimal policy π^* of Lemma 3.1 starting at discontinuity d.*

Proof. Let $(x^{(1)}, i^{(1)}), (x^{(2)}, i^{(2)}), \ldots$ denote the trajectory of π^* starting at d. We proceed inductively. First consider the base case where $d = \min D$. By Lemma 3.1, $x^{(t)}$ must be monotonically non-increasing and, if $i^{(t)} = C_E$, then $x^{(t)} \in D$. By Lemma 3.2, since there are no discontinuities below D, for all $t > 1$, either $i^{(t)} = C_E$ for all $t \geq 1$ or $i^{(t)} = K$ for all $t \geq 1$. The payoff of the former trajectory is computed in Line 7 and the payoff of the latter is computed in Line 3 of Algorithm 1, and the maximum is taken for V_d^* and Trajs_d^*, tie-breaking in favor of Line 3.

For the inductive step, we fix a $d \in D$ where $d < 0$. By Lemma 3.1, $x^{(t)}$ must be monotonically non-increasing and, if $i^{(t)} = C_E$, then $x^{(t)} \in D$. Consider the set of discontinuities visited by the optimal policy π^*: $\left\{x^{(t)} \mid t \geq 2, x^{(t)} \in D\right\}$.

If this set is empty, i.e. no discontinuities are visited, then by Lemma 3.2, $i^{(t)} = K$ for all $t \geq 1$ and Line 3 computes the payoff of π^*. If it is not, Line 3 computes the payoff of a policy π that does not visit any discontinuities.

If this set consists only of d, then $i^{(t)} = C_E$ for all $t \geq 1$ and line 7 computes the payoff of π^*. If it does not, Line 7 computes the payoff of a policy π that does stay at d.

If the smallest element of the set is $d' = \min\left\{x^{(t)} \mid t \geq 2, x^{(t)} \in D\right\}$, suppose that t' is the first timestep where $x^{(t')} = d'$. By Lemma 3.1, $i^{(t)} = \min\{K, x_{t-1} - x^* + C_E\}$ for all $t < t'$. Thus, $J_d(\pi^*) = \widehat{v} + \widehat{\gamma} J_{d'}(\pi^*)$ where $\widehat{v}, \widehat{\gamma}, \text{Traj} \leftarrow \text{GetPayoff}(d, d')$. Since $d' \leq d$ and $x^{(t')} \geq x^{(t'+1)}$, we have by inductive hypothesis that $J_{d'}(\pi^*) = V_{d'}^*$. Thus, Line 6 computes the payoff of π^*. If the smallest element is not d', Line 7 computes the payoff of some policy π that does not visit any discontinuities after d before visiting d'.

In any of the three possible cases, one of Line 3, Line 6 or Line 7 must have computed $J_{d'}(\pi^*)$. Moreover, they will do so before computing the payoffs of any other optimal policies. Noting that Line 3, Line 6, and Line 7 only ever compute the payoffs of valid policies, the optimality of π^* implies that Trajs_d^* describes the action trajectory unrolled by π^*. □

Fact 3.5. *For every discontinuity $d \geq 0$ of f, if $\pi^*(d) \leq C_E$, then Trajs_d^* describes the trajectory of the optimal policy π^* of Lemma 3.1 starting at discontinuity d.*

Proof. Let $(x^{(1)}, i^{(1)}), (x^{(2)}, i^{(2)}), \ldots$ denote the trajectory of π^* starting at d. We proceed inductively. First consider the base case where $d = \max D$. By Lemma 3.1, $x^{(t)}$ must be monotonically non-increasing and, if $i^{(t)} = C_E$, then $x^{(t)} \in D$. Thus, $i^{(t)} = C_E$ for all $t \geq 1$, the payoff of which is computed in Line 9 of Algorithm 1.

For the inductive step, we fix a $d \in D$ where $d > 0$. By Lemma 3.1, $x^{(t)}$ must be monotonically non-decreasing and, if $i^{(t)} = C_E$, then $x^{(t)} \in D$. Consider the set of discontinuities visited by the optimal policy π^*: $\left\{x^{(t)} \mid t \geq 2, x^{(t)} \in D\right\}$.

This set cannot be empty, as by Lemma 3.1, some discontinuity must be reached by π^* in finite time. If this set consists only of d, then $i^{(t)} = C_E$ for all $t \geq 1$ and line 9 computes the payoff of π^*. If it does not, Line 9 computes the payoff of a policy π that does stay at d.

If the largest element of the set is $d' = \max\left\{x^{(t)} \mid t \geq 2, x^{(t)} \in D\right\}$, suppose that t' is the first timestep where $x^{(t')} = d'$. By Lemma 3.1, action $i_t = \begin{cases} \delta & \delta > 0 \wedge t = 1 \\ -K & \text{otherwise} \end{cases}$ where $\delta = d' - d \bmod K + C_E$ for all $t < t'$. Thus, $J_d(\pi^*) = \widehat{v} + \widehat{\gamma} J_{d'}(\pi^*)$ where $\widehat{v}, \widehat{\gamma}, \text{Traj} \leftarrow \text{GetPayoff}(d, d')$. Since $d' \geq d$ and $x^{(t')} \geq x^{(t'+1)}$, we have by inductive hypothesis that $J_{d'}(\pi^*) = V_{d'}^*$. Thus, Line 12 computes the payoff of π^*. If the smallest element is not d', Line 12 computes the payoff of some policy π that does not visit any discontinuities after d before visiting d'.

In either of these possible cases, either Line 9 or Line 12 must have computed $J_{d'}(\pi^*)$. Moreover, they will do so before computing the payoffs of any other optimal policies. Noting that Line 9 and Line 12 only ever compute the payoffs of valid policies, the optimality of π^* implies that Trajs^*_d describes the action trajectory unrolled by π^*. □

If the first action taken by π^* keeps user state constant, i.e. $i^{(1)} = C_E$, then after Line 7, $V_0^* = J(\pi^*)$. If the first action taken by π^* decreases user state, i.e. $i^{(1)} > C_E$, then after Line 7, $V_0^* = J(\pi^*)$. If neither of these are the case, observe that there is some policy π for which $V_0^* = J(\pi)$. If the first action taken by π^* increases user state, i.e. $i^{(1)} < C_E$, then after Line 12, $V_0^* = J(\pi^*)$. Because π^* is optimal, we therefore have as desired that $V_0^* = \max_{\pi \in \Pi} J(\pi)$ and Traj_0^* is the trajectory unrolled by π^*. This proves the correctness of Algorithm 1.

We now assert the runtime of Algorithm 1. Each call to Algorithm 2 incurs a runtime of $O(1) \cdot |x^* - x| / (K - C_E)$. Since we only call Algorithm 2 with discontinuities and treat the gap between discontinuities as a constant, calls made to Algorithm 2 by Algorithm 1 each incurs $O(1)$ runtime. The outer loops of Algorithm 1 run for $k+1$ iterations, while the inner loop runs for up to k iterations. This gives the claimed runtime of $O(k^2)$. □

3.2 Online Learning of Content Selection Policies

We now consider an online learning setting where an app receives a series of users one at a time and must fix a content selection policy for each. Upon receiving a user, whose demand function is unknown to the app and may be selected by an adversary, the app fixes a content selection policy for the user. A trajectory of interactions is then unrolled for a long period of $T \to \infty$ timesteps, during which the app observes nothing. The app is only provided with the final episodic payoff for the user, which it uses to select a new content selection policy for the subsequent user. We do not make assumptions about the complexity of the user demand functions, only that they converge to constant values below $x = -m$ and above $x = m$.

We first consider when an app selects from three content choices that respectively increase user engagement, decrease user engagement, or have no effect on the user. In this setting, our Lemma 3.1 constrains the space of possible optimal policies to a small finite set such that the classical bandit algorithm Exp3 [1] provides a sublinear regret bound.

Theorem 3.6. *Consider a linear content selection problem in an online learning setting where users arrive with unknown demand functions $f_1, \ldots, f_T$ that converge to constant values outside $[-m, m]$. Suppose that the set of available content is restricted to $\mathcal{I} = \{-K, C_E, K\}$. There is an online learning algorithm that, observing only the final episodic payoff for each user, chooses simple policies $\pi_1, \ldots, \pi_T$ such that with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} J_{f_t}(\pi_t) \geq \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O\left(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}\right).$$

Here, $J_{f_t}(\pi)$ denotes the payoff of policy π on a user with demand function f_t and Π denotes the set of simple policies.

Proof. First consider any sequence of demand functions $f_1, \ldots, f_T$. We can consider their rounded counterparts $f'_1, \ldots, f'_T$ defined as

$$f'_t(x) \coloneqq \begin{cases} f'_t(x) = f_t(\lfloor x/(K - C_E) \rfloor (K - C_E)) & x \leq 0 \\ f'_t(x) = f_t(\lfloor x/(K + C_E) \rfloor (K + C_E)) & x \geq 0. \end{cases}$$

This effectively transforms each demand function into a piecewise constant function. Note that, since each demand function f_t takes constant values above m and below $-m$, we know that the discontinuities of its rounded counterpart f'_t lie in the set $[-m, m]$. We also note that each demand function is lower-bounded by its rounded counterpart, i.e., $f'_t(x) \leq f_t(x)$ for all $x \in \mathbb{R}$. That is, the rounded demand function f'_t corresponds to a more difficult user who is less likely to engage and therefore provides less value to the app. Moreover, we can verify that optimal payoff is not affected by this rounding.

As part of our characterization of optimal policies in Lemma 3.1 (more specifically, Lemma B.11), we know all user state trajectories are either stationary, monotonically increasing, or monotonically decreasing. Due to the restricted set of available content, there is only a single action that increases user state and only a single action that decreases user state. Hence, every optimal policy must unroll an action trajectory $i^{(1)}, i^{(2)}, \ldots$ where one of the following holds:

1. the app perpetually decreases user state: $i^{(t)} = K$ for all $t \in \mathbb{Z}_+$,
2. user state remains constant: $i^{(t)} = C_E$ for all $t \in \mathbb{Z}_+$,
3. the app decreases user state then keeps it constant: $i^{(t)} = K$ for $t < T$ and $i^{(t)} = C_E$ for $t > T$ for some finite T,
4. the app increases user state then keeps it constant: $i^{(t)} = -K$ for $t < T$ and $i^{(t)} = C_E$ for $t > T$ for some finite T.

In the first three cases, the user state trajectory $x^{(1)}, x^{(2)}, \ldots$ satisfies, at all timesteps $t \in \mathbb{Z}_+$:

$$x^{(t)} \in \{\lfloor x/(K - C_E) \rfloor (K - C_E) \mid x \in \mathbb{Z}_-\}. \tag{2}$$

In the fourth case, the user state trajectory satisfies:

$$x^{(t)} \in \{\lfloor x/(K + C_E) \rfloor (K + C_E) \mid x \in \mathbb{Z}_+\}. \tag{3}$$

Observe that, due to (2) and (3), any user state $x^{(t)}$ reached by an optimal policy must satisfy $f'_t(x^{(t)}) = f_t(x^{(t)})$. Hence, rounding demand functions has no impact on optimal payoff:

$$\max_{\pi^* \in \widetilde{\Pi}} \sum_{t=1}^{T} J_{f_t}(\pi^*) = \max_{\pi^* \in \widetilde{\Pi}} \sum_{t=1}^{T} J_{f'_t}(\pi^*).$$

The advantage of working with rounded demand functions is that they are piecewise constant with discontinuities spaced out strategically.

Fact 3.7. *Given a set of rounded demand functions $f'_1, \ldots, f'_T$, consider the set of simple policies $\Pi' = \{\pi_{i,v} \mid v \in \pm 1, i \in [0, \ldots, 2m/K] \cup \{-\infty\}\}$ where $\pi_{i,v}(i \cdot K - m) = C_E$ and $\pi_{i,v}(x) = vK$ for all $x \neq i \cdot K - m$. There is an optimal policy in this set, i.e. $\Pi' \cap \arg\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$.*

Fact 3.7 gives that

$$\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) \leq \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f_t}(\pi^*),$$

for the restricted function class Π' as defined in Fact 3.7:

$$\Pi' = \{\pi_{i,v} \mid v \in \pm 1, i \in [0, \ldots, 2m/K] \cup \{-\infty\}\}$$

where $\pi_{i,v}(i \cdot K - m) = C_E$ and $\pi_{i,v}(x) = vK$ for all $x \neq i \cdot K - m$.

Since our reduced policy space Π' from Fact 3.7 is small with $|\Pi'| \in O(\frac{m}{K})$, we can directly apply a bandit learning algorithm to choose from Π'. The only remaining challenge is that, with each user, we only observe a single realization of our policy on the user; hence, the episodic payoff that we receive as feedback is noisy. However, since episodic payoffs are bounded by the geometric series $\frac{1}{1-\gamma}(K + C_R)$, these noisy estimates are not only unbiased but also bounded in $[0, \frac{1}{1-\gamma}(K + C_R)]$.

We can thus apply a standard stochastic approximation argument [12] to the high-probability regret bound of Exp3-IX [13].

Lemma 3.8. *Consider a protagonist who repeatedly uses the (random) Exp3-IX algorithm [13] to select an action i_t from a finite set $\mathcal{A}$. An adversary, who observes $i_1, \ldots, i_t$, chooses a loss $\ell_t \in [0,1]^{\mathcal{A}}$, of which the protagonist only receives a noisy unbiased observation $\widehat{\ell}_t \in [0,1]^{\mathcal{A}}$ where $\mathbb{E}[\widehat{\ell}_t] = \ell_t$. The protagonist then chooses its next action $i_{t+1} \sim \text{Exp3}(\{(i_\tau, \widehat{\ell}_\tau(i_\tau))\}_{\tau \in [t]})$ With probability at least $1 - \delta$ in the randomness of Exp3 and the observed losses $\widehat{\ell}$:*

$$\sum_{t=1}^{T} \widehat{\ell}_t(i_t) \geq \max_{i^* \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(i^*) - O(\sqrt{T\,|A| \log(|A|\,/\delta)}).$$

Applying Exp3-IX results in the regret bound

$$\begin{aligned}
\sum_{t=1}^{T} \widehat{J}_{f_t}(\pi_t) &\geq \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}) \\
&\geq \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f'_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}) \\
&= \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}).
\end{aligned}$$

□

We next consider the online learning setting where an app chooses from a spectrum of content choices. Here, we again apply Lemma 3.1 to constrain the space of approximately optimal policies.

Theorem 3.9. *Consider a linear content selection problem in an online learning setting where users arrive with unknown demand functions $f_1, \ldots, f_T$ that converge to constant values outside $[-m, m]$. There is an online learning algorithm that, observing only a realization of episodic reward after each user, chooses simple policies $\pi_1, \ldots, \pi_T$ such that with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} J_{f_t}(\pi_t) \geq \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O\left(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}\right) - T(K + C_E).$$

If every user has a high initial level of engagement, i.e. $f_t(0) \geq \frac{1}{\gamma}(1 - \frac{C_E + C_R}{2(C_E + K)})$ for all $t \in [T]$:

$$\sum_{t=1}^{T} J_{f_t}(\pi_t) \geq \tfrac{1}{2} \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O\left(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}\right).$$

Proof. First consider any sequence of demand functions $f_1, \ldots, f_T$. We can again consider their rounded counterparts $f'_1, \ldots, f'_T$ defined as

$$f'_t(x) \coloneqq \begin{cases} f'_t(x) = f_t(\lfloor x/(K - C_E)\rfloor (K - C_E)) & x \le 0 \\ f'_t(x) = f_t(\lfloor x/(K + C_E)\rfloor (K + C_E)) & x \ge 0. \end{cases}$$

Recall that the rounded demand function f'_t corresponds to a more difficult user who is less likely to engage and therefore provides less value to the app. We can verify that optimal payoff is not significantly affected by this rounding.

Fact 3.10. *The optimal payoff that can be realized with the rounded demand functions is within $O(T(K + C_E))$ that of the original demand functions:*

$$\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) \le T(K + C_E) + \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f'_t}(\pi^*).$$

Fact 3.10 and Fact 3.7 give

$$\begin{aligned} \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) &\le T(K + C_E) + \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f'_t}(\pi^*) \\ &\le T(K + C_E) + \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f_t}(\pi^*), \end{aligned}$$

where the second inequality applies that $f'_t(x) \le f_t(x)$ for all $x \in \mathbb{R}$.

As in our proof of Theorem 3.6, we can again directly apply a bandit learning algorithm to choose from Π'. We can thus apply a standard stochastic approximation argument [12] to the high-probability regret bound of Exp3-IX [13] on the reduced policy space Π', which results in

$$\begin{aligned} \sum_{t=1}^{T} \widehat{J}_{f_t}(\pi_t) &\ge \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}) \\ &\ge \max_{\pi^* \in \Pi'} \sum_{t=1}^{T} J_{f'_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}) \\ &\ge \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) - O(\sqrt{\tfrac{Tm\gamma}{K(1-\gamma)}(K + C_R)\log(m/K\delta)}) - T(K + C_E). \end{aligned}$$

For the latter claim, observe that the policy π keeping user state constant at 0, i.e. playing $i_t = C_E + C_R$ for all $t \in [T]$, results in a payoff of

$$\frac{1}{1 - \gamma \widetilde{f}(x_0)}(C_E + C_R) \ge \frac{1}{1 - \gamma f(x_0)}(C_E + C_R) \ge 2(K + C_E).$$

Thus, we can change the linear regret term $T(K+C_E) \le \frac{1}{2} \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$ into an approximate regret bound. □

4 Analyzing User Disengagement with Modified Demand Elasticity

In this section, we introduce *modified demand elasticity* as a key primitive that encapsulates how a user's demand function, the friction coefficient, and the discount factor affect an app's optimal content selection policy. We demonstrate that modified demand elasticity is a useful tool for analyzing how changes in our model's basic building blocks, such as friction, impact the optimal policy. In particular, we use modified demand elasticity to uncover a counterintuitive effect of friction on app behaviors: friction can increase user engagement while decreasing revenue. We also show that greater modified demand elasticity implies better alignment between an app creator and their user's satisfaction.

4.1 Modified Demand Elasticity

Demand elasticity is a classical notion of how sensitive consumer demand is to a firm's actions (e.g., prices) and plays a central role in how firms behave: if user demand is sensitive to changes in the firm's actions, firms are incentivized to increase demand. In repeated app-user interactions, the natural analogue of demand elasticity is the sensitivity of user demand to the content shown by an app. We propose the following formula for what we refer to as *modified demand elasticity*:

$$\frac{\partial}{\partial x}\log \widetilde{f}(x) = \frac{\partial}{\partial x}\log\left(f(x) + \frac{1-f(x)}{1-\gamma(1-(1-c)f(x))}(1-c)f(x)\gamma\right).$$

This quantity characterizes how sensitive the adjusted demand function $\widetilde{f}$ (from Theorem 2.2) is to the content shown by an app as reflected in the user state x. Moreover, modified demand elasticity provides sufficient information about a user's demand function f and the friction parameter c to determine an app creator's utility. This can be observed from Theorem 2.2, which provides an equivalent form of the app's objective value that has no explicit dependence on the friction parameter c, except through the definition of $\widetilde{f}$.

4.2 User Engagement is Not Monotone in Friction

It is not hard to see that the optimal payoff that can be realized by an app is monotonically non-increasing in the amount of user friction in the app's content selection problem, i.e. in the parameter c. However, under the optimal policy, the relationship between the amount of user engagement that the app receives is in fact not monotone in the amount of user friction. That is, increases in friction—i.e., decreases in the probability that users return to an app—can result in higher user engagement.

A comparative statics analysis. This phenomenon arises due to the strategic incentives of app creators, which we can analyze with comparative statics. Consider a simple instance of our model where (1) every piece of content $i \in \mathcal{I} \subseteq \mathbb{R}$ has a deterministic effect on user experiences, (2) the derivative of the demand function f is positive everywhere i.e. $\frac{\partial}{\partial i}f(x) > 0$ for all $x \in \mathbb{R}$, and (3) the app creator's optimal policy begins by showing some specific content $i = \mathbf{i}$ in the first interaction and holds the user's state steady thereafter by showing content i' so that $x_2 = x_3 = \cdots$. The app's investment in user engagement can be summarized by this initial choice of content $\mathbf{i}$: if the app's optimal choice of $\mathbf{i}$ is increasing in the model's friction parameter c, then friction is incentivizing the app creator to show content that is conducive to user engagement.

We can write the content selection problem as a Markov decision process (Fact A.2) where the process state is the user's state x_t. As such, we can define a Q-function $Q : \mathcal{I} \times \mathbb{R} \to \mathbb{R}$ where

$$Q(i, x) = \mathbb{E}\left[R_i + \gamma \widetilde{f}(x + E_i) \max_{i \in \mathcal{I}} Q(i, x + E_i)\right]$$

denotes the optimal discounted payoff that the app creator can obtain if it takes an action i when a user is in state x. Since the app creator's policy is optimal, it must be that $\mathbf{i} \in \arg\max_i Q(i, 0)$, so the Q-function has a local maximum at the point $(\mathbf{i}, 0)$, i.e., $\frac{\partial}{\partial i} Q(\mathbf{i}, 0) = 0$ and $\frac{\partial^2}{\partial i^2} Q(\mathbf{i}, 0) < 0$. We will assume that Q is doubly differentiable and that $i \mapsto R_i$ is differentiable. We will write the Q-function with the subscript Q_c to make the dependence on our model's friction parameter c explicit.

We can then use the implicit function theorem [17] to study how the solutions to the first-order condition $\frac{d}{di} Q_c(\mathbf{i}, 0) = 0$ evolve with friction. In particular, it establishes the following relationship between the endogenous variable $\mathbf{i}$ and exogenous parameter c.

$$\frac{\partial \mathbf{i}}{\partial c} = -\left[\frac{\partial^2}{\partial c \partial i} Q_c(\mathbf{i}, 0)\right]\left[\frac{\partial^2}{\partial i^2} Q_c(\mathbf{i}, 0)\right]^{-1}.$$

The inverted term is negative by assumption; thus, we have that $\frac{\partial \mathbf{i}}{\partial c}$ has the same sign as the cross-partial derivative. That is, if the cross-partial derivative is positive, then friction increases the app creator's incentive to invest in engagement. Next, using Theorem 2.2 we can write the Q-function as

$$Q_c(\mathbf{i}, 0) = R_i + \gamma \widetilde{f}(E_\mathbf{i}) \sum_{t=1}^{\infty} (\gamma \widetilde{f}(E_\mathbf{i}))^{t-1} R_{i'} = R_\mathbf{i} + R_{i'} \tfrac{\gamma \widetilde{f}_c(E_\mathbf{i})}{1 - \gamma \widetilde{f}_c(E_\mathbf{i})}.$$

Since $\widetilde{f}(E_\mathbf{i})$ is a function of c and $f(E_\mathbf{i})$, simple algebra shows that

$$\frac{\partial^2}{\partial c \partial i} Q_c(\mathbf{i}, 0) = \frac{\partial^2}{\partial c \partial i}\left(R_\mathbf{i} + R_{i'} \tfrac{\gamma \widetilde{f}_c(E_\mathbf{i})}{1 - \gamma \widetilde{f}_c(E_\mathbf{i})}\right) = \gamma^2 R_{i'} \cdot \frac{\partial f(E_\mathbf{i})}{\partial i} \cdot \underbrace{\frac{(2 - c\gamma) f(E_\mathbf{i}) - 1}{(1 - \gamma)(1 - f(E_\mathbf{i})\gamma c)^3}}_{A}.$$

Here, we used the fact that the derivative $\frac{\partial}{\partial c} R_i$ is zero since an arm's payoff is independent of friction c. For the second term, we simply expanded the definition of $\widetilde{f}(E_\mathbf{i})$ in terms of c and $f(E_\mathbf{i})$ (see Theorem 2.2) and computed the derivative. Since all other terms are strictly positive, A must have the same sign as $\frac{\partial^2}{\partial c \partial i} Q_c(\mathbf{i}, x)$. Therefore, if we have that $(2 - c\gamma) f(E_i) \geq 1$ (for example if $c\gamma = 0.5$ and $f(E_i) \geq 2/3$), then term A is positive. This inequality holds when the app's optimal policy already guarantees a relatively high probability of user engagement. In such a case, a marginal increase in friction incentivizes the app creator to show a piece of content that attracts higher levels of user engagement. We can next consider a concrete example of when this occurs.

Example 4.1. Suppose a user's satisfaction with an app can be categorized according to three levels:

- They dislike the app if their cumulative satisfaction is below a threshold $a \in \mathbb{R}$. In this case, the user will stop interacting with the app.
- They moderately enjoy the app if their cumulative satisfaction falls within an interval $[a, b)$. In this case, the user has a 60% chance of continuing to use the app.
- They are enthusiastic about the app if their cumulative satisfaction is larger than b. In this case, the user has a 99% chance of continuing to use the app.

Further suppose the app creator has a $\gamma = 0.9$ discount factor and a linear content landscape parameterized by $i \in [-b, b]$ such that displaying content i yields $R_i = 1 + i$ revenue for the app and $E_i = -i$ user experience. If we compare the optimal policy of the app creator and the resulting user experience when increasing friction parameter c, we can see that higher friction can result in the app providing content with a strictly better user experience.

Proposition 4.2. In Example 4.1, for an appropriate choice of a, b, the user is strictly less satisfied by the optimal app policy when there is less friction compared to when there is more friction. Formally, for any $c' > c$, let x_t^c and $x_t^{c'}$ denote the user experience states at time step t in the optimal policies for friction parameters c and c'. Then, for all t, $x_t^c \leq x_t^{c'}$ and this inequality is strict for $t \geq 2$.

The role of modified demand elasticity. Modified demand elasticity provides a direct explanation for why friction can incentivize app creators to invest more in user engagement. In particular, $\frac{\partial}{\partial x} \log \widetilde{f}$ is always increasing in friction c. This means that as user friction increases, user demand becomes more sensitive to changes in user state and, by extension, choices in app content. Moreover, the ratio of modified demand elasticity $\frac{\partial}{\partial x} \log \widetilde{f}$, when comparing the setting of complete friction ($c = 1$) and no friction ($c = 0$), is linearly increasing in user demand $f(x)$. This predicts that the increase in modified demand elasticity that results from an increased amount of friction is exaggerated when user demand is high.

A high-engagement regime for friction. Intuitively, if user engagement is already low, an app creator is unlikely to be able to drive user engagement to a high enough level that friction presents a less pressing issue. We can formalize this intuition by revisiting our construction in Example 4.1. To simplify our discussion, we will compare the setting with complete friction ($c = 1$) and no friction ($c = 0$), although our construction easily extends for general increases in friction.

Example 4.3 (Generalization of Example 4.1)**.** Suppose a user's satisfaction with an app can be categorized according to the following three levels:

- If their cumulative satisfaction is below a threshold a, they will stop interacting with the app.
- If their cumulative satisfaction falls within an interval (a, b), they have a p_1 chance of continuing to use the app.
- If their cumulative satisfaction is larger than b, they have a p_2 chance of continuing to use the app.

We consider an app creator who has a γ discount factor and, as in Example 4.1, a linear content landscape parameterized by $i \in [-b, b]$ such that displaying content i yields $R_i = 1 + i$ revenue for the app and $E_i = -i$ user experience. We again compare the app creator's optimal policy and its effect on the user's experience when there is no friction versus complete friction.

Proposition 4.4. In Example 4.3, there is an appropriate choice of a and b such that the user is strictly less satisfied when there is no friction ($c = 0$) than when there is complete friction ($c = 1$). To hold, p_1 and p_2 must satisfy the following criteria:

1. p_1, p_2 are not too far apart: $p_2 - p_1 < \min\left\{\frac{1-\gamma}{\gamma(1-\gamma p_1)}, \frac{1-\gamma}{2\gamma(1-\gamma p_2)}\right\}$,
2. $h(p_2) > h(p_1)$ where $h(p) \coloneqq \frac{1}{1-\gamma p} - \frac{\gamma}{1-\gamma} p$ is visualized in Figure 1.

Proof. This proof will follow similarly to that of Proposition 4.2. Let us choose $a = 0$ and $b = \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1)$ for some sufficiently small choice of $\varepsilon > 0$. We can write our user demand function as

$$f(x) = \begin{cases} 0 & x < 0 \\ p_1 & 0 \leq x < \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1) \\ p_2 & \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1) \leq x. \end{cases}$$

By Lemma 3.1, we know that the optimal app policy must be one of three:

1. Maximally decrease user state at each interaction by repeatedly showing content $i_t = b$.
2. Show content $i_t = 0$ in perpetuity to maintain the user state at $x = 0$.
3. Show content $i_1 = -b$ and subsequently show content $i_t = 0$ in perpetuity to maintain the user state at $x = b$.

As before, the first policy results in the user immediately disengaging and hence a utility of

$$1 + b = 1 + \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1).$$

When there is no friction, the payoff of the second policy is $\frac{\gamma}{1-\gamma}p_1 + 1$ and the payoff of the third policy is $\frac{\gamma}{1-\gamma}(p_1 - \varepsilon) + 1$. By construction, the app prefers the second policy over the third policy.

When there is complete friction, the payoff of the second policy is $\frac{1}{1-\gamma p_1}$ and the payoff of the third policy is $\frac{1}{1-\gamma p_2} - \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1)$. Recall that by criterion (2), we have

$$\frac{1}{1-\gamma p_2} - \frac{\gamma}{1-\gamma}p_2 > \frac{1}{1-\gamma p_1} - \frac{\gamma}{1-\gamma}p_1.$$

Re-arranging, we have that the third policy's payoff has a larger payoff than the second policy:

$$\frac{1}{1-\gamma p_2} - \frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1) > \frac{1}{1-\gamma p_1},$$

where we choose $\varepsilon > 0$ here to be sufficiently small.

It remains only to show that the first policy is always suboptimal. Recall the assumption that $p_2 - p_1$ is not too large, satisfying $p_2 - p_1 < \frac{1-\gamma}{\gamma(1-\gamma p_1)}$ and $p_2 - p_1 < \frac{1-\gamma}{2\gamma(1-\gamma p_2)}$. Let ε be sufficiently small that $p_2 - p_1 \leq \frac{1-\gamma}{\gamma(1-\gamma p_1)} - \varepsilon$ and $p_2 - p_1 \leq \frac{1-\gamma}{2\gamma(1-\gamma p_2)} - \varepsilon$. Since we can therefore compute

$$\frac{1}{1-\gamma p_2} > 1 + 2\frac{\gamma}{1-\gamma}(\varepsilon + p_2 - p_1),$$

the payoff of the third policy under complete friction always exceeds the friction-independent payoff of the first policy. □

The second criterion of Proposition 4.4 is usually the binding one. Figure 1 illustrates the function h. To read whether condition 2 of Proposition 4.4 is satisfied using Figure 1, find the point (p_1, y_1) on the graph of $h(p)$ corresponding to the x-value of p_1 and find the point (p_2, y_2) on the graph corresponding to the x-value of p_2; if the y-value corresponding to p_2 is larger, i.e. $y_2 > y_1$, then the criterion is satisfied. Observe that the graph of Figure 1 looks like the letter U. This means that condition 2 of Proposition 4.4 holds—and thus, friction incentivizes the app to increase engagement—if p_1, p_2 are both close to 1, because h is increasing in this region. Meanwhile,

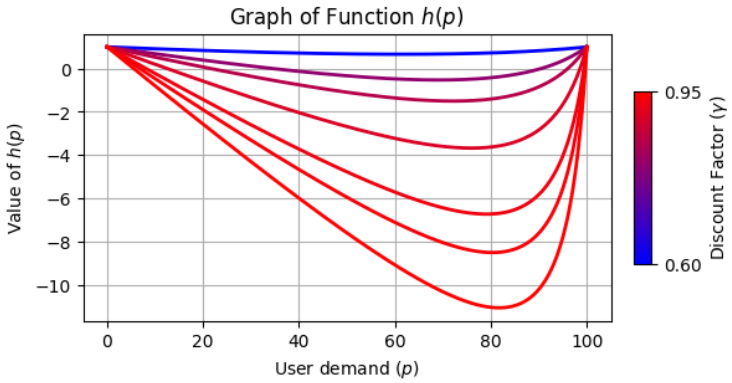

Figure 1: Plot of the function $h(p) \coloneqq \frac{1}{1-\gamma p} - \frac{\gamma}{1-\gamma} p$ on the domain $p \in [0, 1]$ for various choices of γ.

condition 2 does not hold if p_1 and p_2 are both small (for example, if p_2 is smaller than the minimum of the curve corresponding to the app's discount factor γ).

This function h is the difference between modifying the geometric series $\sum_{t=1}^{\infty} \gamma^{t-1}$ by 1) scaling γ down by p versus 2) scaling down the entire series by $p\gamma$. Intuitively, scaling γ down by p reflects the impact that disengagement has on the app creator's utility under complete friction: in each timestep, there is a $1 - p$ chance that future revenue completely disappears. Meanwhile, scaling the entire series by $p\gamma$ reflects the impact of disengagement when there is no friction: payoff in each timestep is scaled down by a constant factor to reflect the constant probability of disengagement. The event that $h(p_2)$ is larger than $h(p_1)$ tells us that, when user demand is at p_2, the app creator's utility is more affected by increasing friction from $c = 0$ to $c = 1$.

4.3 Attributing Friction Phenomena to Modified Demand Elasticity

Under the optimal app policy, we can expect that the user's state will quickly reach an equilibrium level x_∞, as formalized in Lemma 3.1. We can use modified demand elasticity to analyze how friction impacts x_∞. In order to hold the user state steady at this equilibrium, the app will repeatedly show the user a specific type of content. Given a user state x, let $U(x, c)$ be the asymptotic utility that the app derives from repeatedly showing the user content $i \in \mathcal{I}$ when they are in state x. Formally,

$$U(x, c) \coloneqq \frac{\mathbb{E}[R_i]}{1 - \gamma \widetilde{f}(x)},$$

where $\widetilde{f}$ is the modified demand function as discussed before and defined in Theorem 2.2. Said another way, $U(x, c)$ approximates the utility that awaits an app if it quickly drives a user's state to an equilibrium $x_\infty = x$. The difference in asymptotic app utility at different user states x, x', $U(x, c) - U(x', c)$, tells us roughly how much an app creator should be willing to sacrifice in revenue to change a user state from x' to x.

Figure 2 plots asymptotic app utility at different levels of user demand $f(x)$ and friction c. When user demand is high, i.e. $f(x) \to 1$, asymptotic app utility grows at a faster rate, with steepness increasing in friction. To interpret this steepness, consider the following example. Since the difference between asymptotic app utility at demands $f(x) = 0.9$ and $f(x) = 0.8$ is much larger when friction is $c = 1$ than when friction is smaller at $c = 0.5$, the app creator is willing to sacrifice more revenue to increase user demand (from $0.8 \to 0.9$) when friction is higher.

To reason about the curvature of asymptotic app utility more formally, let us consider its

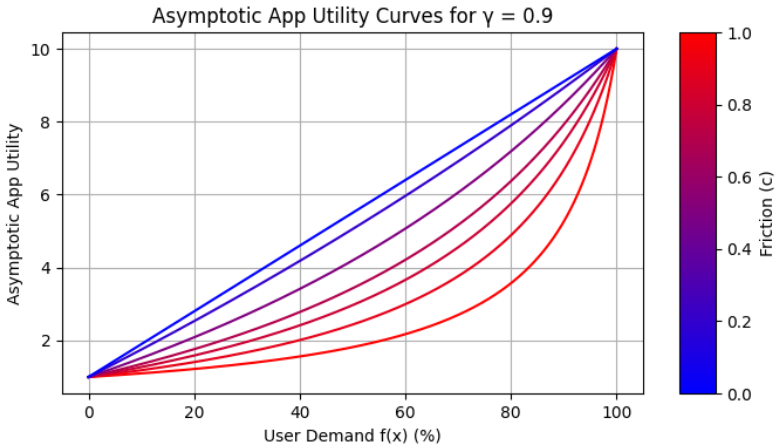

Figure 2: Asymptotic app utility at different levels of user demand and friction. Content revenue and the app creator's discount factor are fixed at $\mathbb{E}[R_i] = 1$ and $\gamma = 0.9$ for clarity.

derivative, which we can decompose into three interpretable factors:

$$\frac{d}{dx}\frac{1}{1-\gamma\widetilde{f}(x)} = \underbrace{\left(\frac{1}{1-\gamma\widetilde{f}(x)}\right)^2}_{(A)} \cdot \underbrace{\gamma\widetilde{f}(x)}_{(B)} \cdot \underbrace{\frac{d}{dx}\log(\widetilde{f}(x))}_{(C)}, \tag{4}$$

Term A is just asymptotic app utility, squared. Term B is the app creator's "effective discount factor", as defined in Section 2.1. Term C is our definition of modified demand elasticity. Whereas terms A and B are strictly decreasing with friction, term C–modified demand elasticity–strictly increases in friction. That is, the fact that *modified demand elasticity* increases with friction is the sole mechanism behind friction-driven increases to user engagement. This decomposition also offers an answer for why we observe that an increase in friction can drive an increase in user engagement when user demand is high, i.e. when $f(x)$ is large. Figure 3 plots the ratios of each of these terms when there is complete friction ($c = 1$) against when there is no friction ($c = 0$). Whereas the asymptotic app utility ratio is symmetric in user demand around $f(x) = 0.5$, the ratios of the effective discount factor and asymptotic app utility are exactly linearly growing in demand.

4.4 Comparison to Classical Demand Elasticity

Our model is a generalization of the classical supply-demand curve to a setting where demand manifests over repeated interactions. Whereas the classical demand function maps the prices chosen by a firm to consumer demand, our model's demand function maps the accumulated effect of the content chosen by an app to user demand. Given this similarity, it may be tempting to consider the classical notion of demand elasticity, which is defined as $\frac{\partial}{\partial x}\log f$ where f is the demand function. In comparison to modified demand elasticity, this quantity $\frac{\partial}{\partial x}\log f$, which we will refer to as classical demand elasticity, does not take into account friction or the app creator's discount factor.

We can gain some direct intuition for why accounting for friction is necessary in defining demand elasticity by plotting the ratio between modified demand elasticity and its classical definition, as is done in Figure 4. This ratio forms a convex curve, with curvature that decreases in friction c and increases in the app creator's discount factor γ. That is, when there is less friction and apps are more patient, modified demand elasticity is significantly smaller than its classical counterpart.

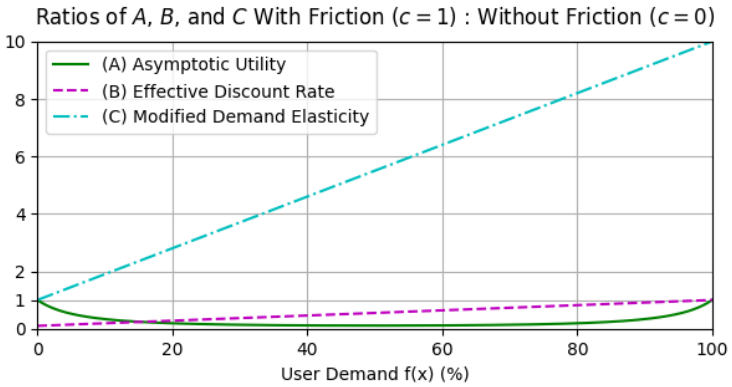

Figure 3: The ratio of the factors (A), (B), and (C) that compose $\frac{\partial}{\partial x}\frac{1}{1-\gamma\widetilde{f}(x)}$ as stated in (4) when there is full friction ($c = 1$) against when there is no friction ($c = 0$). The app creator's discount factor is fixed at $\gamma = 0.9$.

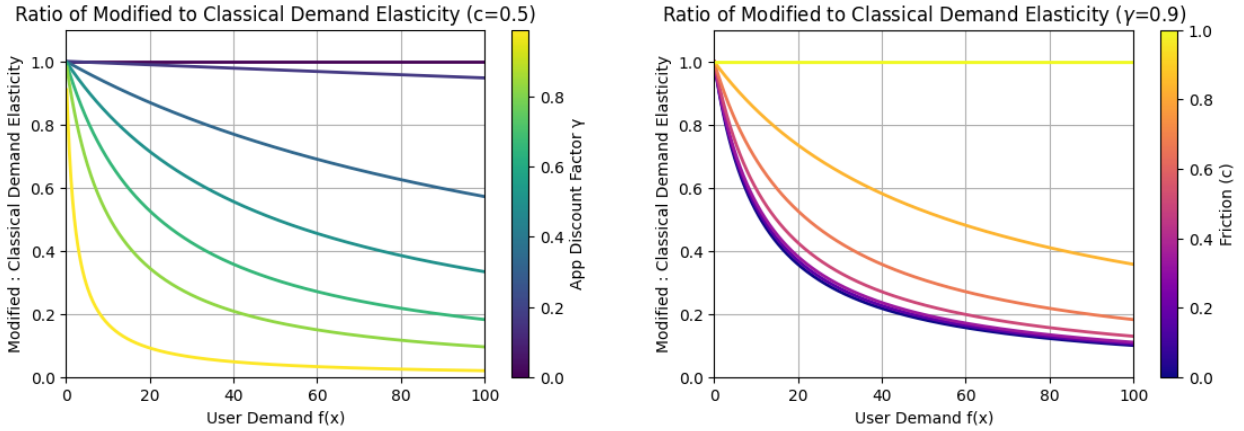

Figure 4: Ratio of modified demand elasticity $\frac{d}{dx}\log\widetilde{f}(x)$ to classical demand elasticity $\frac{d}{dx}\log f(x)$.

This curvature captures an important but simple intuition: when the app creator's discount factor γ is large and friction c is small, app creators are less worried about friction keeping users away as they can afford to patiently wait for users to return if they disengage. In contrast, higher levels of friction amplify even small perturbations to user demand, an amplification that means investments in increasing user demand can become increasingly profitable.

This also clarifies why modified demand elasticity appears as the central quantity in our model: an app creator's behavior depends on how they perceive user demand rather than literal user demand. In contrast, these two concepts coincide in classical supply-demand curves. We also emphasize that the gap between an app creator's perception of user demand, $\widetilde{f}(x)$, and literal user demand, $f(x)$, does not arise due to the information structure of our model. That is, the app creator's perception of user demand is not affected by a lack of information about user demand or model parameters; rather, the distinction between perceived and literal demand arises from the repeated nature of interactions in our model and the necessary corrections that app creators must make to optimize for long-term utility.

4.5 Demand Elasticity and Alignment

Modified demand elasticity also provides a foundation on which to prove statements about app behavior in our model for general settings. For example, we can prove an asymptotic lower bound on how much user demand an app should aim to foster. We will analyze a form of partial alignment between user welfare and the app's revenue, which quantifies the revenue a user must provide an app to incentivize the app to augment the user's welfare.

Theorem 4.5 states that, given any x^* and δ, if the lifetime customer value exceeds a certain amount x^*, the app's optimal policy should always invest in increasing the user's state so that it remains above the threshold x^* for at least $1-\delta$ of the time. Importantly, the lifetime customer value needed is *decreasing* in the modified demand elasticity of the user. Intuitively, while the value of a user is the "reward" available to an app by increasing engagement, the modified demand elasticity of the user is inversely related to the "cost" of increasing engagement.

Theorem 4.5. *Consider any threshold in user state $x^* \in \mathbb{R}$ and small constant δ. Suppose that an app sees in its user an achievable payoff of at least*

$$\max_{\pi} J(\pi) \in \Omega\left(\frac{\log(1/\gamma)}{\delta} \cdot \left(\frac{\partial}{\partial x} \log \widetilde{f}(x) \mid_{x=x^*}\right)^{-1}\right).$$

Then, any app policy where the user's state is below x^* for at least a δ-fraction of interactions—that is, $\frac{1}{T}\sum_{t=1}^{T} 1[x^{(t)} \leq x^*] \geq \delta$ for all large T—is suboptimal. Here, Ω treats content attributes $\{E_i\}_{i\in\mathcal{I}}$ and $\{R_i\}_{i\in\mathcal{I}}$ as constants.

5 Discussion

This paper develops a model for the algorithmic problem of content selection, with the goal of capturing the rich multi-objective nature of trading off between maximizing immediate revenues and increasing future user engagement. As a starting point, and to highlight the tractability of the model, we show that, under mild linearity assumptions, optimal policies are well-structured and can be efficiently computed. Moreover, we demonstrate that—because of the structure that optimal policies possess in our model—there always exists a small approximate covering of the policy space such that simple online learning guarantees can even be shown with out-of-the-box bandit algorithms. This paper also applied our model as a microfoundation of recommendation systems, to understand how the primitives of content selection affect the alignment between recommendation systems and the users they serve. We identified modified demand elasticity to be the key primitive affecting whether a platform is incentivized to select content that is higher engagement or content that is higher revenue. We used this primitive to demonstrate that making it harder for users to interact with an app they are not currently using (i.e. increasing friction) may counter-intuitively boost user-app engagement.

There are several directions for future work. While this paper provided initial results on learning optimal content selection policies from data, the online learning guarantees (Theorem 3.9 and Theorem 3.6) presented are not tight and naively apply out-of-the-box algorithms—improved guarantees are achievable with more careful analysis and algorithms. Although this paper proved a sufficient condition for app-user alignment (Theorem C.3), proving tight alignment guarantees remains an open problem. Cleanly characterizing when apps are incentivized to invest in user engagement would provide important insight into the strategic behavior of, for example, the recommendation engines of social media platforms. An open question also remains around when the counter-intuitive phenomenon we observed with friction (Section 4.2) arises in competitive multi-app settings.

6 Acknowledgements

This work was supported in part by the National Science Foundation under grants CCF-2145898 and CCF-2338226, by the Office of Naval Research under grant N00014-24-1-2159, a C3.AI Digital Transformation Institute grant, the Mathematical Data Science program of the Office of Naval Research, and Alfred P. Sloan fellowship, and a Schmidt Science AI2050 fellowship. This material is based upon work also supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Calvano acknowledges financial support from the ERC-ADV grant 101098332 and PRIN 2022 CUP E53D23006420001.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[2] M. Baucells and R. K. Sarin. Satiation in Discounted Utility. *Operations Research*, 55(1): 170–181, Feb. 2007. ISSN 0030-364X. doi: 10.1287/opre.1060.0322. Publisher: INFORMS.

[3] O. Ben-Porat, L. Cohen, L. Leqi, Z. C. Lipton, and Y. Mansour. Modeling Attrition in Recommender Systems with Departing Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6072–6079, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i6.20554. Number: 6.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1st edition, 1995. ISBN 1886529124.

[5] J. Cao, W. Sun, Zuo-Jun, Shen, and M. Ettl. Fatigue-aware Bandits for Dependent Click Models, Aug. 2020. arXiv:2008.09733 [cs, stat].

[6] M. Friedman. The marshallian demand curve. *Journal of Political Economy*, 57(6):463–495, 1949.

[7] J. Kleinberg, S. Mullainathan, and M. Raghavan. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization, June 2022. arXiv:2202.11776 [cs].

[8] R. Kleinberg and N. Immorlica. Recharging Bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319, Oct. 2018. doi: 10.1109/FOCS.2018.00037. ISSN: 2575-8454.

[9] P. Laforgue, G. Clerici, N. Cesa-Bianchi, and R. Gilad-Bachrach. A Last Switch Dependent Analysis of Satiation and Seasonality in Bandits, Mar. 2022. arXiv:2110.11819 [cs].

[10] L. Leqi, F. Kilinc Karzan, Z. Lipton, and A. Montgomery. Rebounding bandits for modeling satiation effects. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[11] A. Marshall. *Principles of economics: unabridged eighth edition*. Cosimo, Inc., 2009.

[12] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[13] G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3168–3176, 2015.

[14] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1): 50–69, 2010.

[15] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang. Stochastic Bandits with Linear Constraints. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, Mar. 2021. ISSN: 2640-3498.

[16] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

[17] D. M. Topkis. *Supermodularity and complementarity*. Princeton university press, 1998.

[18] Z. Yang, X. Liu, and L. Ying. Exploration, Exploitation, and Engagement in Multi-Armed Bandits with Abandonment, May 2022. arXiv:2205.13566 [cs].

[19] H. Zhang, Y. Cheng, and V. Conitzer. Efficient algorithms for planning with participation constraints. In D. M. Pennock, I. Segal, and S. Seuken, editors, *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pages 1121–1140. ACM, 2022. doi: 10.1145/3490486.3538280.

A Omitted Proofs and Additional Results for Section 2

A.1 Proof of Lemma 2.1

Lemma 2.1 is a consequence of the fact that most reasonable Markov decision processes admit an optimal policy that is deterministic and stationary.

Lemma 2.1. *If there is an optimal policy for the app creator, there is also a simple optimal policy.*

Proof. We will first prove that discounted payoff is well-defined (Fact A.1). We then write our model as a Markov decision process and demonstrate a bijection between simple policies of our model and stationary policies of the MDP (Fact A.2). Finally, we recall that there always exists an optimal stationary policy for an MDP with an optimal policy (Fact A.3).

We first can verify that discounted payoff (1) is well-defined (see, e.g., Proposition A.7).

Fact A.1. *For any app policy π, its discounted payoff $J(\pi)$ as defined in* (1) *is finite for valid discount factors $\gamma \in (0, 1)$.*

Proof of Fact A.1. Since we assume that the set of content revenue means $\{\mathbb{E}[R_i]\}_{i\in\mathcal{I}}$ is a compact set, there exists a constant $K \in \mathbb{R}$ that upper bounds $|\mathbb{E}[R_i]| \le K$ for every content $i \in I$. Thus, we can upper bound the payoff by the series $|J(\pi)| \le \sum_{t=1}^{\infty} \gamma^{t-1} |r_t| \le \frac{K}{1-\gamma}$. □

We next observe that there always exists a stationary Markov decision process (MDP) that is equivalent to our model. Fix a user with demand function f and friction c, and an app with content $\mathcal{I}$ that returns revenues $\{R_i\}_{i\in\mathcal{I}}$ and user experiences $\{E_i\}_{i\in\mathcal{I}}$. Let us construct an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ with state space $\mathcal{S} \coloneqq \mathbb{R} \times \{0, 1\}$, actions $\mathcal{A} \coloneqq \mathcal{I} \cup \{\emptyset\}$, transition function $\mathcal{P}$ and reward function $\mathcal{R}$. In the following, we will use S_t to denote the MDP state and a_t to denote the MDP action at timestep t.

At every state in the set $\mathbb{R} \times \{0\} \subset \mathcal{S}$, only the actions in the set $\mathcal{I}$ are available. Each of these states, $(x, 0)$, corresponds to a timestep in which the user is interacting with the app and has a user state of x. At every state in the set $\mathbb{R} \times \{1\} \subset \mathcal{S}$, only a trivial action $\emptyset$ is available. Each of these states, $(x, 1)$, corresponds to a timestep where the user declined to interact with the app. We accordingly define the MDP's initial state to be $(0, 0)$. We define the reward function as $\mathcal{R}(a = \emptyset, S = (x, 1)) = 0$ and $\mathcal{R}(a = i, S = (x, 0)) = R_i$ for all $x \in \mathbb{R}, i \in \mathcal{I}$.

We now define transition probabilities. First, for all $x \in \mathbb{R}$, we define $\mathbb{P}(S_{t+1} = (x, 0) \mid a_t = \emptyset, S_t = (x, 1)) = (1 - c)f(x)$ and $\mathbb{P}(S_{t+1} = (x, 1) \mid a_t = \emptyset, S_t = (x, 1)) = 1 - (1 - c)f(x)$ according to the probability that a user returns to an app after having left it. Similarly, for all $x, y \in \mathbb{R}$ and $i \in \mathcal{I}$, we define $\mathbb{P}(S_{t+1} = (y, 0) \mid a_t = i, S_t = (x, 0)) = \Pr(E_i = y - x)f(y)$ and $\mathbb{P}(S_{t+1} = (y, 1) \mid a_t = i, S_t = (x, 0)) = \Pr(E_i = y - x)(1 - f(y))$ according to the probability that a user continues use of an app. If E_i is continuously valued, we can instead define the cumulative density function $\mathbb{P}(S_{t+1} \in \{(y', 0), y' \le y\} \mid a_t = i, S_t = (x, 0)) = \Pr(E_i \le y - x)\,\mathbb{E}[f(y') \mid y' \le y]$ and $P(S_{t+1} \in \{(y', 1), y' \le y\} \mid a_t = i, S_t = (x, 0)) = \Pr(E_i \le y - x)\,\mathbb{E}[1 - f(y') \mid y' \le y]$.

There is a bijection between MDP policies and app policies that preserves their discounted payoff.

Fact A.2. *There exists a bijective mapping ϕ from the MDP's policy space, i.e. functions of form $\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^* \to \mathcal{A}$, to the space of app policies in our model, i.e. functions of form $\pi : \mathbb{R}^4 \to \mathcal{I}$. This map ϕ guarantees that for every MDP policy π, the MDP's discounted payoff J' matches that of its counterpart in our model: $J'(\pi) = J(\phi(\pi))$. ϕ also maps stationary MDP policies only to simple app policies in our model.*

Proof of Fact A.2. Fix a policy for the MDP: $\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^* \to \mathcal{A}$. We can express π as taking a transcript as input, which we can denote as $H = [(S_1, a_1, r_1), \ldots, (S_{T-1}, a_{T-1}, r_{T-1})] \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^*$. We now define a mapping ψ of MDP transcripts to transcripts in our model. Fix any MDP transcript H, and define $(x_{t+1}, 1 - s_t) = S_t$ for every $t \in [T]$. We construct our mapping ψ as $\psi(H) = \{(s_t, r_t, x_{t+1} - x_t, a_t)\}_{t \in [T-1]}$.

We can define our bijection ϕ as $\phi(\pi) = \pi(\psi^{-1}(\cdot))$; in other words, for every possible transcript of user-app interactions H, the policy $\phi(\pi)$ returns the action $\pi(\psi^{-1}(H))$. Writing J' to denote the discounted objective of the MDP and using J as in (1), by construction, we have $J'(\phi(\pi)) = J(\pi)$ for every policy π. Similarly, we have $J(\phi^{-1}(\pi)) = J'(\pi)$ for every MDP policy π.

We can also confirm that stationary policies in the MDP indeed map to simple app policies. This is because every policy in the MDP only plays non-trivial actions, i.e., $a_t \neq \emptyset$, at states that belong in the set $\mathbb{R} \times \{0\}$. In those states, since a stationary policy's action at a timestep t depends only on the current state $S_{t-1} = (x_t, 0)$, the stationary policy's action—and by extension its corresponding app policy's action—only depends on the user's state x_t. □

We thus have that the existence of an optimal app policy for in our model implies the existence of an optimal policy for the MDP. Moreover, any deterministic and stationary policy for the MDP implies a simple app policy attaining the same objective value. The following statement concerning policy iteration concludes our proof.

Fact A.3 (Bertsekas [4])**.** *In a stationary MDP, for any (potentially history dependent and non-deterministic) policy, there exists a stationary and deterministic policy with at least as high an expected discounted payoff, on every initial state distribution.*

□

A.2 Existence of Simple Optimal Policies

An optimal app policy can be shown to exist under standard and mild assumptions inherited from the theory of Markov decision processes.

Proposition A.4. There is an optimal simple app policy in our model if at least one of the following conditions hold:

1. All contents provide deterministic user experiences, i.e. $\{E_i\}_{i \in \mathcal{I}}$ are deterministic.
2. The app chooses from a finite set of content.
3. The set of app content is compact and the CDF of content user experiences, $\Pr(E_i \leq z)$, is continuous in the content $i \in \mathcal{I}$.

Proof. The MDP is trivial on states in the set $\mathbb{R} \times \{1\}$. It therefore suffices for us to define a Bellman operator exclusively on the state space $\mathbb{R} \times \{0\}$, which we will map onto the reals for convenience. For any function $W : \mathbb{R} \to \mathbb{R}$, we define the Bellman operator T as

$$(TW)(x) = \max_{i \in \mathcal{I}} \mathop{\mathbb{E}}_{e \sim E_i} \left[\mathop{\mathbb{E}}_{\substack{n \sim N(x,e) \\ r \sim \mathcal{R}(a=i, S=(x,0))}} [r + \gamma^n W(x+e)] \right],$$

where we define $N(x, y)$ as the random variable that is the number of timesteps that the MDP takes to reach the state $(x + y, 0)$ from the state $(x, 0)$, conditioned on the MDP reaching either the state $(x + y, 0)$ or $(x + y, 1)$ immediately after state $(x, 0)$. Fact A.5 and Fact A.6 prove two important properties of this operator.

Fact A.5. *Under any of the listed assumptions, $TW(x)$ is well-defined for every monotonically non-decreasing W and for every $x \in \mathbb{R}$.*

Proof. In Case 1, when user experiences are deterministic, the operator can be written as

$$(TW)(x) = \max_{i \in \mathcal{I}} \mathop{\mathbb{E}}_{\substack{n \sim N(x,E_i) \\ r \sim \mathcal{R}(a=i,S=(x,0))}} \left[r + \gamma^n W(x + E_i) \right],$$

where we note that $N(x, E_i)$ is monotonically non-increasing in E_i as the user demand functions are monotonically non-decreasing and γ^n is monotonically non-increasing in n as $\gamma < 1$. We also have that $W(x + E_i)$ is monotonically non-decreasing in E_i by assumption. It thus follows that the expectation inside the maximum, namely

$$\mathop{\mathbb{E}}_{\substack{n \sim N(x,E_i) \\ r \sim \mathcal{R}(a=i,S=(x,0))}} \left[r + \gamma^n W(x + E_i) \right],$$

is non-decreasing in E_i. Since $\{E_i\}_{i \in \mathcal{I}}$ is compact, the maximum exists and hence $TW(x)$ exists.

In Case 2, it is also obvious that $TW(x)$ exists when contents are finite.

In Case 3, the below expectation is necessarily continuous

$$g(i) = \mathop{\mathbb{E}}_{e \sim E_i} \left[\mathop{\mathbb{E}}_{\substack{n \sim N(x,e) \\ r \sim \mathcal{R}(a=i,S=(x,0))}} \left[r + \gamma^n W(x + e) \right] \right].$$

Since $\mathcal{I}$ is a compact set, the existence of $TW(x)$ follows by the extreme value theorem. □

Fact A.6. *If W is monotonically non-decreasing and bounded below by zero, then TW is also monotonically non-decreasing and bounded below by zero.*

Proof. By assumption $f(x)$ is monotonically non-decreasing in x. Thus, $N(x, E_i)$ is weakly stochastically dominated by $N(x', E_i)$ if $x > x'$. Since γ^n is non-increasing in n and the function $W \geq 0$ is bounded below by zero for all arguments, it also follows that TW is monotonically non-decreasing. Since there always exists a content with non-negative revenue, i.e. $\exists i \in I$ such that $\mathbb{E}[R_i] \geq 0$, it is also true that TW remains bounded below by zero. □

By the previous facts, if we repeatedly apply the operator T to a monotonically non-decreasing function W bounded below by zero, say the all zeros function $W(x) = 0$, we will always obtain a well-defined monotonically non-decreasing function bounded below by zero. A standard argument of the γ-contractiveness of T and appeal to Brouwer's fixed point theorem [16] then directly implies the existence of a monotonically non-decreasing non-negative function W^* such that $TW^* = W^*$. There must therefore exist an optimal stationary policy π defined as

$$\pi(x) = \arg\max_{i \in \mathcal{I}} \mathop{\mathbb{E}}_{e \sim E_i} \left[\mathop{\mathbb{E}}_{\substack{n \sim N(x,e) \\ r \sim \mathcal{R}(a=i,S=(x,0))}} \left[r + \gamma^n W(x + e) \right] \right].$$

By Lemma 2.1, there therefore also exists a simple app policy that is optimal. □

A.3 Non-Existence of Average Reward Objective

A common alternative to studying discounted reward objectives is studying the average reward objective. In our model, however, the average reward objective may not be well-defined for some stationary policies.

Proposition A.7. There exists an instance of our model in which a simple policy does not have a well-defined expected average reward:

$$\lim_{T\to\infty} \mathop{\mathbb{E}}_{\{s_t,r_t,e_t,i_t\}_t} \left[\frac{1}{T}\sum_{t=1}^{T} r_t\right].$$

Proof. Consider an instance of our model where the app creator chooses from two pieces of content $I = \{a, b\}$, each of which give deterministic user experiences $E_a = E_b = 1$ and revenues $R_a = -1$, $R_b = 1$. Let $f(0) = 1$ for simplicity. Consider the following simple policy $\pi : \mathbb{R} \to \{a, b\}$ where for each $q \in \mathbb{Z}$ and every $x \in (3^q, 3^{q+1}]$, we set $\pi(x) = a$ if q is odd and $\pi(x) = b$ if q is even. That is, the policy alternates between showing content a for increasing long periods to showing content b for increasing long periods, such that the empirical content distribution has no limit.

The deterministic trajectory of this policy will be user states $1, 2, \ldots$ and revenues

$$-1, +1, +1, -1, -1, -1, -1, -1, -1, +1, \ldots.$$

Direct computation gives that the average reward series alternates between values of -0.5 and 0.5 with $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_t \le -0.5$ and $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_t \ge 0.5$. The limit that is expected average reward therefore does not exist. □

B Omitted Proofs and Additional Results for Section 3

B.1 Greedy Revenue Optimization

Consider a user that already has a user state x_t that is at an extreme value, whether it is extremely high (user demand is near maximum) or extremely low (user demand is near minimum). We will show that app creators are always incentivized to act greedily with such users and show the highest revenue-earning content.

Proposition B.1. Suppose the user experiences provided by app content are of finite variance. For any $\varepsilon > 0$, there is a threshold $x^* \in \mathbb{R}$ where for all higher user states $x^{(t)} \ge x^*$, the app creator has an ε-optimal simple policy that shows the highest revenue earning content in the next timestep, i.e. $i^{(t+1)} = \arg\max_{i\in\mathcal{I}} \mathbb{E}[R_i]$. Similarly, there exists a threshold $x^* \in \mathbb{R}$ where for all lower user states $x^{(t)} \le x^*$, the app creator is again ε-incentivized to show the highest revenue earning content.

Proof. We want to show that for any $\varepsilon > 0$, for sufficiently large or small x,

$$Q(\arg\max_{i\in\mathcal{I}} \mathbb{E}[R_i], x) \ge \max_{i\in\mathcal{I}} Q(i, x) - \varepsilon.$$

First, we recall that, by the compactness of the revenue expectations $\{\mathbb{E}[R_i]\}_{i\in\mathcal{I}}$, we can upper bound the amount of revenue that the app creator can attain in a single interaction by some constant value $K \in \mathbb{R}$. It therefore follows that the value function at any user state $x \in \mathbb{R}$ is upper bounded by the geometric series $V(x) \le \frac{K\gamma}{1-\gamma}$. We also know the value function is non-negative and monotonically non-decreasing at all user states $x \in \mathbb{R}$. Since the domain of the value function V is $\mathbb{R}$ yet the range

of V is bounded, we know that for every $\delta > 0$ and $\varepsilon > 0$ there exists a threshold x^* such that for all higher user states $x > x^*$, we have

$$V(x) - V(x - \delta) \leq \varepsilon.$$

For convenience, we now define $i_{\text{greedy}} \coloneqq \arg\max_{i \in \mathcal{I}} \mathbb{E}[R_i]$ to be the content that greedily maximizes expected revenue, which must exist by the compactness of $\{\mathbb{E}[R_i]\}_{i \in \mathcal{I}}$. We now similarly define $i_{\text{opt}} \coloneqq \arg\max_{i \in \mathcal{I}} Q(i, x)$ to be the content that the app shows if it continues executing some optimal policy. We can write the difference in the Q-function values of showing the optimal content i_{opt} and showing the greedy content i_{greedy} as

$$\begin{aligned}
&Q(i_{\text{opt}}, x) - Q(i_{\text{greedy}}, x) \\
&= \widetilde{f}(x)\gamma \left(\mathbb{E}\left[R_{i_{\text{opt}}}\right] - \mathbb{E}\left[R_{i_{\text{greedy}}}\right] + \mathbb{E}\left[V(x + E_{i_{\text{opt}}})\right] - \mathbb{E}\left[V(x + E_{i_{\text{greedy}}})\right] \right) \\
&\leq \widetilde{f}(x)\gamma \left(\mathbb{E}\left[V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - (E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}))\right] \right) \\
&\leq \mathbb{E}\left[V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - (E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}))\right].
\end{aligned}$$

In the final inequality, we use the fact that $Q(i_{\text{opt}}, x) \geq Q(i_{\text{greedy}}, x)$ by definition of i_{opt} and the fact that $\widetilde{f}(x) \leq 1$ and $\gamma \leq 1$.

We now use the fact that $E_{i_{\text{opt}}}$ and $E_{i_{\text{greedy}}}$ each have bounded variance and finite means to invoke Chebyshev's inequality. Chebyshev's inequality gives that, for any probability $p \in (0, 1)$, there exists a δ' such that $\left|E_{i_{\text{opt}}}\right| \leq \delta'$ with probability at least $1 - p/2$. We can similarly apply Chebyshev's inequality to the random variable $E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}$, which is also of bounded variance and finite mean. By Chebyshev's inequality, for any $p > 0$, there exists a δ such that $\left|E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}\right| \leq \delta$ with probability at least $1 - p/2$. Suppose that, for some fixed choice of δ, δ', we condition on the event $\left|E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}\right| \leq \delta$ and $\left|E_{i_{\text{opt}}}\right| \leq \delta'$. It then immediately follows that there exists some user state x^* such that for all greater user states $x > x^*$,

$$V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - (E_{i_{\text{opt}}} - E_{i_{\text{greedy}}})) \leq V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - \delta) \leq \varepsilon/2.$$

Applying a union bound, we know that there is some choice of δ, δ' such that the aforementioned event $\left|E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}\right| \leq \delta$ and $\left|E_{i_{\text{opt}}}\right| \leq \delta'$ occurs with probability at least $1 - p$. When this event does not occur, we can still apply a deterministic upper bound to the value function of

$$V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - (E_{i_{\text{opt}}} - E_{i_{\text{greedy}}})) \leq \frac{K\gamma}{1 - \gamma}.$$

Thus, if we choose the probability $p = \frac{\varepsilon}{2}\frac{1-\gamma}{K\gamma}$, we have that there exists a user state x^* such that for all $x \geq x^*$

$$\begin{aligned}
Q(i_{\text{opt}}, x) - Q(i_{\text{greedy}}, x) &\leq \mathbb{E}\left[V(x + E_{i_{\text{opt}}}) - V(x + E_{i_{\text{opt}}} - (E_{i_{\text{opt}}} - E_{i_{\text{greedy}}}))\right] \\
&\leq (1 - p)\frac{\varepsilon}{2} + p\frac{K\gamma}{1 - \gamma} \\
&\leq \varepsilon,
\end{aligned}$$

where the second inequality applies the law of total expectation.

The second half of the claim—the existence of a threshold x^* such that for all lower user states $x \leq x^*$ apps are also at least ε-incentivized to show the highest revenue content—follows identically. □

B.2 Proof of Lemma 3.1

Lemma 3.1. *In the linear setting where the user demand function has a complexity of $k < \infty$, there is an optimal simple policy for the app creator that satisfies all of the following characteristics:*

- *The sequence $x_1, x_2, \ldots$ of user states is monotonic.*
- *The limit $x_\infty = \lim_{t\to\infty} x_t$ exists and is either at a discontinuity of f or negative infinity.*
- *(When $x_\infty = -\infty$) The app always shows the highest-revenue content, i.e. $i = K$.*
- *(When x_∞ is a discontinuity of f) The user state x_∞ will be reached within $k + n + 1$ interactions (i.e., $x_{n+k+1} = x_\infty$) where n is the smallest number of interactions in which x_∞ can be reached in some policy (i.e., there exists a policy where $x_n = x_\infty$). Moreover, if $n = 1$, $x_2 = x_3 = \cdots = x_\infty$.*

Proof. We refer to the sequence of user states $x^{(1)}, x^{(2)}, \ldots$ that result from executing a simple policy π as the *user state trajectory* of π. In the linear setting, this trajectory is deterministic. We similarly refer to the sequence of content $i^{(1)}, i^{(2)}, \ldots$ that result from executing a simple policy π as the *action trajectory* of π. We will use the shorthand D to denote the set of discontinuities of the demand function f. Our proof consists of three steps. We will first establish that the set of content maximizing the Q-function at each state is compact. We then accordingly construct a specific instance of an optimal simple policy π^*. We conclude by proving our four claims.

We begin by remarking that, since the demand function f is piecewise linear and right-continuous with finite discontinuities, $\widetilde{f}$ is also piecewise linear and right-continuous with finite discontinuities.

Fact B.2. *The function $\widetilde{f}$ is piecewise linear and right-continuous with $|D| < \infty$ discontinuities.*

We next observe that any simple optimal policy should only "hold" user state constant at user states that correspond to discontinuities of the user demand function f. In the sequel, we will use $V_\pi(x)$ to denote the discounted payoff that results from executing a simply policy π on a user with initial state x. Observe that the value function can be written as $V(x) = \max_\pi V_\pi(x)$. We will also write $Q_\pi(i, x)$ to denote the discounted payoff that results from showing a content i to a user with initial state x and thereafter executing a simple policy π.

Lemma B.3. *A simple optimal policy π should only keep user states constant at discontinuities of f: that is, if for some $x \in \mathbb{R}$ and simple policy π, the equality $\pi(x) = C_E$ holds and π is optimal, then $x \in D$.*

Proof. Fix a user state $x \in \mathbb{R}$ and suppose to the contrary that it is not a discontinuity: $x \notin D$. Since there are only finite discontinuities, there must exist a lower level of user state $x' < x$ such that user engagement is unchanged: $f(x) = f(x')$. Moreover, we can choose x' so that we also have that x' is reachable from x within one timestep: $x' \geq x + C_E - K$. For example, we can define $x' = \max\{d \in D \mid d < x\} \cup \{x + C_E - K\}$. We will use $i' \in (0, K]$ to denote the content that one shows to lower user state from x to x' in a single timestep.

Let π be a simple policy that holds user state constant at x, i.e. set $\pi(x) = C_E$ as $E_{C_E} = 0$. We will design a policy π' to be a witness to the suboptimality of π by attaining a strictly larger discounted objective value at an initial user state of $x^{(0)} = x$. We define this policy π' as first showing the content i' to lower user state from x to x', and then holding user state constant into perpetuity by repeatedly showing the content $i_{\text{hold}} = C_E$. This is a valid construction as $i' \in [-K, K]$ by construction. Using the geometric series identity, we can explicitly write

$$V_\pi(x) = \frac{C_R + C_E}{1 - \widetilde{f}(x)\gamma}, \quad V_{\pi'}(x) = \frac{C_R + C_E}{1 - \widetilde{f}(x)\gamma} + x - x',$$

where we use the fact that $\widetilde{f}(x') = \widetilde{f}(x)$. Since $x > x'$ by construction, we have as desired that $V_{\pi'}(x) > V_\pi(x)$ and a witness to the suboptimality of π. □

The next lemma constrains the possible trajectories of simple optimal policies when the user state trajectory does not coincide with the discontinuities D.

Lemma 3.2. *Consider an optimal simple policy π. For any subsequence of the policy's user state trajectory, $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$, where $x^{(2)}, \ldots, x^{(T)}$ are all not discontinuities of f:*

1. *If the policy π is not maximally increasing user state at the last step of the subsequence, i.e. $\pi(x^{(T)}) > -K$, then all previous actions in the subsequence should be maximally decreasing user state $\pi(x^{(1)}) = \pi(x^{(T-1)}) = K$.*

2. *If the policy π is not maximally decreasing user state at the first step, that is $\pi(x^{(1)}) < K$, then all later actions in the subsequence should be maximally increasing user state $\pi(x^{(2)}) = \pi(x^{(T)}) = -K$.*

Proof. We begin by proving the first claim. Suppose to the contrary that for some intermediate user state $x^{(0)} \in \mathbb{R}$ (which we re-index to timestep 0 for notational convenience), applying the policy π results in a user state trajectory with the subsequence $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$ where the following conditions simultaneously hold:

1. The user states $x^{(2)}, \ldots, x^{(T)}$ do not coincide with any of the discontinuities of f.

2. π does not maximally increase user state at the end of the subsequence: $\pi(x^{(T)}) > -K$.

3. At some $t \in [T-1]$, the app does not maximally decrease user state: $\pi(x^{(t)}) < K$.

There must exist some $\delta > 0$ so that $\pi(x^{(t)}) + \delta \leq K$, $\pi(x^{(T)}) - \delta \geq -K$ and $f(x^{(\tau)} - \delta) = f(x^{(\tau)})$ for all $\tau \in [t-1, T]$. In fact, we can simply choose

$$\delta \coloneqq \min\left\{K - \pi(x^{(t)}), \pi(x^{(T)}) + K\right\} \cup \left\{0.5 \cdot |d - x^{(\tau)}| \mid \tau \in [t+1, T], d \in D\right\}.$$

This δ is bounded away from zero since we assume that $x^{(2)}, \ldots, x^{(T)}$ are not discontinuities of f and well-defined since there are only finite discontinuities in D.

We now construct a policy π' which yields the trajectory $x^{(1)}, \ldots, x^{(t)}, x^{(t+1)} - \delta, \ldots, x^{(T)} - \delta$ and resumes the trajectory of π at the $(T+1)$th timestep. This policy π' is valid because we chose δ so as to ensure that $\pi(x^{(t)}) + \delta \leq K$ and thus $\pi'(x^{(t)}) \leq K$. Similarly, we chose δ so that $\pi(x^{(T)}) - \delta \geq -K$ and by extension $\pi'(x^{(T)}) \geq -K$. Thus, at the only two timesteps in which the actions of policy π differ from π', timesteps t and T, we are still guaranteed the new actions are legal: $\pi'(x^{(t)}) \in [-K, K]$ and $\pi'(x^{(T)}) \in [-K, K]$.

We can verify this policy π' guarantees a strictly higher discounted payoff than π, leading to a contradiction with the optimality of π. Let us write the user state trajectory of the original policy π as $x^{(1)}, x^{(2)}, \ldots$ and that of the modified policy π' as $x^{(1)}, \ldots, x^{(t-1)}, \widehat{x}^{(t)}, \ldots, \widehat{x}^{(t)}, x^{(T+1)}, \ldots$. Let us also write the original actions trajectory of policy π as $i_1, i_2, \ldots$. Then we can express each of

the payoffs of π and π' as

$$\begin{aligned}
V_\pi(x_0) &= \sum_{w=1}^{\infty} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i_w + C_R), \\
V_{\pi'}(x_0) &= \sum_{w=1}^{t-1} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i_w + C_R) + \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(t)} + \delta + C_R) \\
&\quad + \sum_{w=t+1}^{T-1} \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(\prod_{\tau=t+1}^{w} \gamma \widetilde{f}(\widehat{x}^{(\tau)}) \right) (i^{(w)} + C_R) \\
&\quad + \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(\prod_{\tau=t+1}^{T} \gamma \widetilde{f}(\widehat{x}^{(\tau)}) \right) (i^{(T)} - \delta + C_R) \\
&\quad + \sum_{w=T+1}^{\infty} \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(\prod_{\tau=t+1}^{T} \gamma \widetilde{f}(\widehat{x}^{(\tau)}) \right) \left(\prod_{\tau=T+1}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R).
\end{aligned}$$

Since we chose δ to guarantee that $\widetilde{f}(\widehat{x}^{(\tau)}) = \widetilde{f}(x^{(\tau)})$ at all timesteps $\tau \in [t, T]$, we can thus simplify the payoff of policy π' as

$$\begin{aligned}
V_{\pi'}(x_0) &= \sum_{w=1}^{\infty} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R) + \delta \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(1 - \left(\prod_{\tau \in [t+1,T]} \gamma \widetilde{f}(x^{(\tau)}) \right) \right) \\
&= V_\pi(x_0) + \delta \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(1 - \left(\prod_{\tau \in [t+1,T]} \gamma \widetilde{f}(x^{(\tau)}) \right) \right) \\
&> V_\pi(x_0).
\end{aligned}$$

Thus, π' is thus a witness to the suboptimality of π.

We can very similarly prove the second claim. Suppose to the contrary that for some choice of initial user state, applying the policy π results in a user state trajectory with the initial subsequence $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$ where the following conditions simultaneously hold:

1. The user states $x^{(2)}, \ldots, x^{(T)}$ do not coincide with any of the discontinuities of f.
2. The policy π does not maximally decrease user state as its first action: $\pi(x^{(1)}) < K$.
3. At some $t \in [2, T]$, the app does not maximally increase user state: $\pi(x^{(t)}) > -K$.

As before, there must exist some small value $\delta > 0$ so that $\pi(x^{(t)}) - \delta \geq -K$, $\pi(x^{(1)}) + \delta \leq K$ and $f(x^{(2)} - \delta) = f(x^{(2)}), \ldots f(x^{(t)} - \delta) = f(x^{(t)})$. This time, we construct a (non-simple) policy π' which executes the trajectory $x^{(1)}, x^{(2)} - \delta, \ldots, x^{(t)} - \delta, x^{(t+1)}, \ldots, x^{(T)}$ and then resumes the trajectory of π starting at the $(t+1)$st timestep. Again, we note that this policy is valid, since we chose δ so as to ensure that $\pi(x^{(1)}) + \delta \leq K$ and $\pi(x^{(t)}) - \delta \geq -K$.

We also observe this policy π' guarantees a strictly higher objective value than π, leading to a contradiction with the optimality of π. Let us write the user state trajectory of the original policy π as $x^{(1)}, x^{(2)}, \ldots$ and that of the modified policy π' as $x^{(1)}, x^{(2)}, \ldots, \widehat{x}^{(t)}, x^{(t+1)}, \ldots$. Let us also write the actions trajectory of policy π as $i^{(1)}, i^{(2)}, \ldots$. Then we can express each of the objectives of π

and π' as

$$\begin{aligned}
V_\pi(x_0) &= \sum_{w=1}^{\infty} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R), \\
V_{\pi'}(x_0) &= i^{(1)} + \delta + C_R \\
&\quad + \sum_{w=2}^{t-1} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R) \\
&\quad + \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(t)} - \delta + C_R) \\
&\quad + \sum_{w=t+1}^{\infty} \left(\prod_{\tau=2}^{t} \gamma \widetilde{f}(x^{(\tau)}) \right) \left(\prod_{\tau=t+1}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R).
\end{aligned}$$

We chose δ so that $f(x^{(\tau)} - \delta) = f(x^{(\tau)})$ for all timesteps $\tau \in [2, t]$, meaning we can simplify the payoff of policy π' as

$$\begin{aligned}
V_{\pi'}(x_0) &= \sum_{w=1}^{\infty} \left(\prod_{\tau=2}^{w} \gamma \widetilde{f}(x^{(\tau)}) \right) (i^{(w)} + C_R) + \delta \left(1 - \left(\prod_{\tau \in [2,t]} \gamma \widetilde{f}(x^{(\tau)}) \right) \right) \\
&= V_\pi(x_0) + \delta \left(1 - \left(\prod_{\tau \in [2,t]} \gamma \widetilde{f}(x^{(\tau)}) \right) \right) \\
&> V_\pi(x_0).
\end{aligned}$$

□

In the next lemma, we confirm that there is no optimal simple policy that can result in an action trajectory that infinitely boosts user state.

Lemma B.4. *If there is a simple policy π and initial user state $x \in \mathbb{R}$ such that the action trajectory of π has a limit of maximally boosting user state, i.e. $\lim_{t \to \infty} i^{(t)} = -K$, π cannot be an optimal policy.*

Proof. By definition, for any $\delta > 0$, there is some finite time T past which, for all timesteps $t \geq T$, the actions played are almost maximally boosting user state: $i^{(t)} \leq -K + \delta$. Since there are finite discontinuities, it follows that past some finite time T', all future timesteps $t > T'$ result in user states $x^{(t)}$ exceeding the largest discontinuity: $x^{(t)} \geq \max D$. We thus have that past the timestep $\max\{T', T\}$, the policy π will always play an action $i^{(t)}$ where $i^{(t)} < C_E$, despite user state already resulting in the maximum engagement probability, i.e. $x^{(t)} \geq \max D$. Observe that $x^{(\max\{T', T\})}$ is a deterministic user state.

We have that the payoff of the policy π with initial user state $x^{(\max\{T', T\})}$ is strictly upper bounded by that of any policy π' which keeps user state constant: $\pi'(x^{(\max\{T', T\})}) = C_E$.

$$\begin{aligned}
V_\pi(x^{(\max\{T', T\})}) &\leq (C_R - K + \delta) \frac{1}{1 - \gamma \widetilde{f}(x^{(\max\{T', T\})})} \\
&< (C_R - K) \frac{1}{1 - \gamma \widetilde{f}(x^{(\max\{T', T\})})} \\
&= V_{\pi'}(x^{(\max\{T', T\})}).
\end{aligned}$$

Thus, π is suboptimal. $\square$

We can use the previous lemmas to then prove our main intermediate step, establishing the compactness of the set of optimal actions at every user state.

Lemma B.5. *At every level of user state $x \in \mathbb{R}$, the set of optimal actions optimizing the Q function is finite. That is,* $\arg\max_{i\in[-K,K]} Q(i,x)$ *is finite.*

Proof. We will fix any user state $x \in \mathbb{R}$ and denote the set of optimal actions at x with the shorthand $S \coloneqq \arg\max_{i\in[-K,K]} Q(i,x)$. We argue that every action $i \in S$ must either maximally increase or decrease user state, i.e. $i = K$ or $i = -K$, or satisfy the following inequality for at least one choice of discontinuity $d \in D$: $0 = |x + C_E + i - d| \pmod K$. Note that since there are finitely many discontinuities, the above assertion directly implies the finiteness of S.

Suppose to the contrary that there is an $i \in S$ that does not satisfy any of the above conditions. Then the user state $x + C_E - i$ cannot be a discontinuity of f and the content $i < K$ cannot be maximally decreasing user state. Our trajectory lemma (Lemma 3.2) therefore says that there must exist an optimal policy π where $\pi(x + C_E - i) = -K$. By recursive application of Lemma 3.2, the action trajectory of the policy π at an initial user state of $x + C_E - i$ must repeat the action $i = -K$ until user state trajectory happens to land on a discontinuity of f. However, since we have assumed that there is no $d \in D$ for which $0 = |x + C_E + i - d| \pmod K$, the user state trajectory will never land on a discontinuity, and the action trajectory must be an infinite repetition of the action $-K$. We reach a contradiction by Lemma B.4, as π therefore cannot be optimal. $\square$

We now let π^* denote the simple optimal policy that, for every user state, chooses the content that maximizes the Q function while tie-breaking in favor of actions that decreases user state, that is $\pi(x) = \max\arg\max_{i\in[-K,K]} Q(i,x)$. Lemma B.5 ensures that π^* is well-defined.

Next, we prove the final technical lemma of this proof, which states that, regardless of which pair of initial user states one chooses, the user state trajectories of π^* will never "cross".

Lemma B.6. *Let $\left\{x^{(t)}\right\}_{t\in\mathbb{N}}$ denote the user state trajectory of the policy π^* starting at an initial user state of $x^{(0)}$ and let $\left\{\widehat{x}^{(t)}\right\}_{t\in\mathbb{N}}$ denote the user state trajectory of π^* starting at an initial user state of $\widehat{x}^{(0)}$. The two trajectories will never cross. That is, if $x^{(0)} < \widehat{x}^{(0)}$, then for all $t \in \mathbb{N}$, $x^{(t)} \le \widehat{x}^{(t)}$, and if $x^{(0)} > \widehat{x}^{(0)}$, then for all $t \in \mathbb{N}$, $x^{(t)} \ge \widehat{x}^{(t)}$.*

Proof. Without loss of generality, assume that $x^{(0)} < \widehat{x}^{(0)}$. Suppose to the contrary that there exists a timestep $t \in \mathbb{N}$ at which $x^{(t)} > \widehat{x}^{(t)}$. There must exist some initial timestep $t \in \mathbb{N}$ where the trajectories crossed; that is, where $x^{(t-1)} \le x^{(t-1)}$ and $x^{(t)} > \widehat{x}^{(t)}$. Since π^* is constructed to be a simple policy, we can directly infer that $x^{(t-1)} < x^{(t-1)}$. For notational convenience and without loss of generality, we will assume $t = 2$.

We proceed by observing that the user state $\widehat{x}^{(2)}$ must be reachable from $x^{(1)}$, i.e. $\widehat{x}^{(2)} \in [C_E - K + \widehat{x}^{(1)}, C_E + K + \widehat{x}^{(1)}]$. To see this, recall that we can explicitly write the action $i^{(1)}$ that the policy π^* takes to get from user state $x^{(1)}$ to $x^{(2)}$ is given by the equality $i^{(1)} = x^{(1)} - x^{(2)} + C_E$. Since $x^{(1)} = x^{(0)} < \widehat{x}^{(0)} = \widehat{x}^{(1)}$, it follows that $x^{(1)} - \widehat{x}^{(2)} - C_E < \widehat{x}^{(1)} - \widehat{x}^{(2)} - C_E \le K$. Similarly, we know that $x^{(1)} - \widehat{x}^{(2)} - C_E \ge -K$; otherwise, we would reach a contradiction due to $\widehat{x}^{(2)} > x^{(1)} + K - C_E$ and $x^{(2)} > \widehat{x}^{(2)}$ implying that $x^{(2)} > x^{(1)} + K - C_E$ despite $x^{(2)}$ being reachable from $x^{(1)}$.

Similarly, the user state $x^{(2)}$ must be reachable from $\widehat{x}^{(1)}$, i.e. $x^{(2)} \in [C_E - K + x^{(1)}, C_E + K + x^{(1)}]$. This is because since $x^{(2)}$ is reachable from $x^{(1)}$, $\widehat{x}^{(1)} - x^{(2)} - C_E > x^{(1)} - x^{(2)} - C_E \ge -K$. Since $\widehat{x}^{(2)}$ is reachable from $\widehat{x}^{(1)}$, $\widehat{x}^{(1)} - x^{(2)} - C_E < \widehat{x}^{(1)} - \widehat{x}^{(2)} - C_E \le K$.

Since $x^{(1)}$ can reach both $x^{(2)}$ and $\widehat{x}^{(2)}$, the optimality of the simple policy π^* gives that

$$C_E + C_R + x^{(1)} - x^{(2)} + \widetilde{f}(x^{(2)})\gamma V(x^{(2)}) \geq C_E + C_R + x^{(1)} - \widehat{x}^{(2)} + \widetilde{f}(\widehat{x}^{(2)})\gamma V(\widehat{x}^{(2)}).$$

The optimality of π^* also implies that

$$C_E + C_R + \widehat{x}^{(1)} - \widehat{x}^{(2)} + \widetilde{f}(\widehat{x}^{(2)})\gamma V(\widehat{x}^{(2)}) \geq C_E + C_R + \widehat{x}^{(1)} - x^{(2)} + \widetilde{f}(x^{(2)})\gamma V(x^{(2)}),$$

Thus, $\widetilde{f}(x^{(2)})\gamma V(x^{(2)}) - x^{(2)} = \widetilde{f}(\widehat{x}^{(2)})\gamma V(\widehat{x}^{(2)}) - \widehat{x}^{(2)}$.

We now compare the optimality of following the action suggested by the policy π^* at the user state $\widehat{x}^{(1)}$ with the optimality of showing the content $i = x^{(2)} - \widehat{x}^{(1)} + C_E$, finding

$$\begin{aligned} Q(\pi^*(\widehat{x}^{(1)}), \widehat{x}^{(1)}) &= \widehat{x}^{(2)} - \widehat{x}^{(1)} + C_E + C_R + \widetilde{f}(\widehat{x}^{(2)})\gamma V(\widehat{x}^{(2)}) \\ &= x^{(2)} - \widehat{x}^{(1)} + C_E + C_R + \widetilde{f}(x^{(2)})\gamma V(x^{(2)}) \\ &= Q(x^{(2)} - \widehat{x}^{(1)} + C_E, \widehat{x}^{(1)}). \end{aligned}$$

This implies that $x^{(2)} - \widehat{x}^{(1)} + C_E \in \arg\max_{i \in [-K,K]} Q(i, \widehat{x}^{(1)})$, which is a contradiction, since $x_2 - \widehat{x}^{(1)} + C_E > \widehat{x}^{(2)} - \widehat{x}^{(1)} + C_E$ but $\pi^*(\widehat{x}^{(1)})$ is defined to be the most exploitative of all of the optimal arms: $\pi^*(\widehat{x}^{(1)}) = \max \arg\max_{i \in [-K,K]} Q(i, \widehat{x}^{(1)})$. □

We now prove the facts that compose Lemma 3.1.

Fact B.7. *Any trajectory of the policy π^* is either monotonically non-decreasing or monotonically non-increasing in the user's state.*

Proof. Suppose the policy π^* unrolls a user state trajectory $x^{(1)}, \ldots, x^{(t)}, \ldots$ where $\pi^*(x^{(t)}) > C_E$ and $\pi^*(x^{(t-1)}) < C_E$ or $\pi^*(x^{(t)}) < C_E$ and $\pi^*(x^{(t-1)}) > C_E$. Then, since π^* is a simple policy, the user state trajectory of π^* on the initial user state $x^{(t-1)}$ crosses the user state trajectory of π^* on the initial user state $x^{(t)}$. This contradicts Lemma B.6. □

Fact B.8. *In any trajectory of the policy π^* where user states increase, there can be at most $k+1$ timesteps in which the step is neither a fixed point, i.e. $i = C_E$, nor a full step upwards, i.e. $i = -K$.*

Proof. Suppose that the action trajectory of the policy π^* simultaneously satisfies $i^{(t)} \in (-K, C_E)$ and $i^{(t')} \in (-K, C_E)$. Without loss of generality, suppose $t < t'$. Lemma 3.2 directly implies that, in at least one timestep, the user state trajectory subsequence $x^{(t+1)}, \ldots, x^{(t')}$ is a discontinuity; that is $\exists \tau \in [t+1, t']$ where $x^{(\tau)} \in D$. Since there are only k discontinuities of the function f, there can only be $k+1$ timesteps with incomplete steps $i \in (-K, C_E)$. □

Fact B.9. *In any trajectory of the policy π^* where user states decrease, there can be at most $k+1$ timesteps in which the step is neither a fixed point, i.e. $i = C_E$, nor a full step downwards, i.e. $i = K$.*

Proof. As in the proof of Fact B.8, suppose that the action trajectory of the policy π^* simultaneously satisfies $i^{(t)} \in (C_E, K)$ and $i^{(t')} \in (C_E, K)$ where $t < t'$. Lemma 3.2 directly implies that, in at least one timestep, the user state trajectory subsequence $x^{(t+1)}, \ldots, x^{(t')}$ is a discontinuity. Since there are only k discontinuities of the function f, there can only be $k+1$ timesteps with incomplete steps $i \in (C_E, K)$. □

Fact B.10. *The action trajectory of the policy π^* always has a limit which exists, i.e. $\lim_{t\to\infty} i^{(t)}$ exists, and the limit must either be maximal exploitation, i.e. $\lim_{t\to\infty} i^{(t)} = K$, or maintaining user state constant. In the latter case, the limit of the user state trajectory also exists and is one of the two neighboring discontinuities d at which π^* keeps user state with $\pi^*(d) = C_E$.*

Proof. Fact B.7 states that the user state trajectory of the policy π^* is monotonically non-decreasing or monotonically non-increasing. Since π^* is a simple policy, once it reaches a user state at which it keeps user state constant, i.e. $i = C_E$, it will do so perpetually. We also know that the policy π^* will only play actions in the intervals (C_E, K) and $(-K, C_E)$ a finite number of times. Thus, the limit of the action trajectory of policy π^* must always either be C_E, $-K$ or K.

We can rule out the limit of the policy being the action that maximally boosts user state, i.e. $\lim_{t\to\infty} i^{(t)} = -K$, by Lemma B.4. In the case that the limit is maintaining user state at a constant level, i.e. $\lim_{t\to\infty} i^{(t)} = C_E$, we recall that the limit must occur on a discontinuity of f. In this case, the user state trajectory must also have a limit, where either $\lim_{t\to\infty} x^{(t)} = \min\left\{d \in D \mid d > x^{(0)}\right\}$ or $\lim_{t\to\infty} x^{(t)} = \max\left\{d \in D \mid d < x^{(0)}\right\}$. This is because we know, by Lemma B.3, that user state can only be kept constant on a discontinuity, i.e. $\lim_{t\to\infty} x^{(t)} \in D$. Moreover, by Lemma B.6, it is impossible for the limiting user state to satisfy $\lim_{t\to\infty} x^{(t)} > \min\left\{d \in D \mid d > x^{(0)}\right\}$ as the trajectory of π^* would cross the trajectory of π^* starting at the user state $\lim_{t\to\infty} x^{(t)}$. Similarly, Lemma B.6 guarantees $\lim_{t\to\infty} x^{(t)} \geq \max\left\{d \in D \mid d < x^{(0)}\right\}$. □

□

We can verify that Lemma 3.2 and Lemma 3.1 trivially extend to optimal policies for multiple users. Lemma B.11 follows identically as Lemma 3.1.

Lemma B.11. *In the linear setting where a set of user demand functions $f_1, \ldots, f_T$ each has a complexity of $k < \infty$, there is an optimal simple policy $\pi \in \arg\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$ for the app creator that satisfies all of the following characteristics:*

- *The sequence $x_1, x_2, \ldots$ of user states is monotonic.*
- *The limit $x_\infty = \lim_{t\to\infty} x_t$ exists and is either at a discontinuity of f or negative infinity.*
- *(When $x_\infty = -\infty$) The app always shows the highest-revenue content, i.e. $i = K$.*
- *(When x_∞ is a discontinuity of f) The user state x_∞ will be reached within $k+n+1$ interactions (in other words $x_{n+k+1} = x_\infty$) where n is the smallest number of interactions in which x_∞ can be reached in some policy (that is, if there exists an app policy for which $x_n = x_\infty$). Moreover, if $n = 1$, $x_2 = x_3 = \cdots = x_\infty$.*

Lemma B.12 follows identically as Lemma 3.2.

Lemma B.12. *Consider a simple policy π that is optimal for a set of user demand functions $f_1, \ldots, f_T$, i.e. $\pi \in \arg\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$. For any subsequence of the policy's user state trajectory, $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$, where $x^{(2)}, \ldots, x^{(T)}$ are all not discontinuities for any demand function in $f_1, \ldots, f_T$:*

1. *If the policy π is not maximally increasing user state at the last step of the subsequence, i.e. $\pi(x^{(T)}) > -K$, then all previous actions in the subsequence should be maximally decreasing user state $\pi(x^{(1)}) = \pi(x^{(T-1)}) = K$.*
2. *If the policy π is not maximally decreasing user state at the first step, that is $\pi(x^{(1)}) < K$, then all later actions in the subsequence should be maximally increasing user state $\pi(x^{(2)}) = \pi(x^{(T)}) = -K$.*

B.3 Proofs of Fact 3.7 and Fact 3.10

Fact 3.7. *Given a set of rounded demand functions $f'_1, \ldots, f'_T$, consider the set of simple policies $\Pi' = \{\pi_{i,v} \mid v \in \pm 1, i \in [0, \ldots, 2m/K] \cup \{-\infty\}\}$ where $\pi_{i,v}(i \cdot K - m) = C_E$ and $\pi_{i,v}(x) = vK$ for all $x \neq i \cdot K - m$. There is an optimal policy in this set, i.e. $\Pi' \cap \arg\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$.*

Proof. Lemma B.11 (Lemma 3.1) establishes that the user state trajectory of an optimal policy must be either stationary, monotonically increasing or monotonically decreasing. We will consider this optimal policy and the three possible types of trajectories it can take, and prove that in all three situations, the optimal policy must exist in Π'.

The stationary case is trivial as there exists a policy $\pi_{1,m/K} \in \Pi'$ that always plays $i^{(t)} = C_E$ and thus keeps user state constant.

Now consider the monotonically decreasing case. Suppose, for the sake of contradiction, that the optimal policy is not always fully decreasing user state, i.e. it plays $i^{(t)} < K$ at some timestep t. This means that $x^{(t)} \bmod (K - C_E) \neq 0$ at some timestep t. Then Lemma B.12 states that the user will keep fully decreasing user state until it reaches a discontinuity of the demand function. Every discontinuity x of the demand function satisfies $x \bmod (K - C_E) = 0$; this means that $x^{(t')} \bmod (K - C_E) = 0$ at some timestep $t' > t$; t' must be finite as all discontinuities of f'_t lie in $[-m, m]$. However, fully decreasing user state decreases user state by $K - C_E$, meaning that there must exist some finite j such that $x^{(t)} - j(K - C_E) \bmod (K - C_E) = x^{(t)} \bmod (K - C_E) = 0$, which is a contradiction. Since the demand function is constant below $-m$ and above m and thus all discontinuities must lie in the set $\{iK - m \mid i \in [0, \ldots, 2m/K]\}$, we have that $\pi_{1,i}$ must be optimal for some choice $i \in [0, \ldots, 2m/K] \cup \{-\infty\}$.

The monotonically increasing case follows similarly. Lemma B.12 gives that the user will fully increase user state until it reaches a discontinuity of the demand function. Fully decreasing user state decreases user state by $K - C_E$. As before, the optimal policy must be fully increase user state until reaching a fixed point, which is represented by $\pi_{-1,i}$ for some i. □

Fact 3.10. *The optimal payoff that can be realized with the rounded demand functions is within $O(T(K + C_E))$ that of the original demand functions:*

$$\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*) \leq T(K + C_E) + \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f'_t}(\pi^*).$$

Proof. Let $\pi^* = \arg\max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*)$ denote the optimal policy for the original demand functions. Note that a user's demand function does not influence the per-interaction trajectory of policy π^*, only the amount of time between each interaction. Thus, its (deterministic) trajectory $(x^{(1)}, i^{(1)}), (x^{(2)}, i^{(2)}), \ldots$ on every user will be the same.

We can now consider the non-stationary but Markov policy π' that executes the action trajectory

$$\widetilde{i}^{(t)} = \max\{-K, i^{(t)} - \max\{0, K + C_E - (\sum_{\tau < t} i^{(\tau)} - \widetilde{i}^{(\tau)})\}\}.$$

That is, π' executes the same action trajectory as the optimal policy π^* except that, initially, π' maximally investments in boosting the user state until the user state is $K + C_E$ what it would have been under the policy π^*. In doing so, π' guarantees that it produces as much demand on the rounded demand functions as π^* does on the original demand functions, i.e.

$$f'_t(\widetilde{x}^{(t)}) \geq f'_t(\left\lceil x^{(t)}/(K + C_E) \right\rceil (K + C_E)) \geq f_t(x^{(t)})$$

for all timesteps t. Moreover, the only difference in the actions taken by π' and π^* is that π' invests $K + C_E$ more into user state. It follows that

$$J_{f_t}(\pi^*) - J_{f_t}(\pi') \leq \sum_{t=1}^{T} i^{(t)} - \widetilde{i}^{(t)} \leq K + C_E.$$

Since there must exist a stationary Markov policy with as high a payoff as π', we can upper bound the left-hand side of our claim by

$$T(K + C_E) + \sum_{t=1}^{T} J_{f_t}(\pi') \leq T(K + C_E) + \max_{\pi^* \in \Pi} \sum_{t=1}^{T} J_{f_t}(\pi^*).$$

□

C Omitted Proofs and Additional Results for Section 4

C.1 Proof of Theorem 4.5

We will first define the following notions for convenience.

Definition C.1. Given a simple policy π and some $\delta > 0$, we say that the user state is (x^*, δ) bounded if there exists a constant T such that for all $T' \geq T$, $\frac{1}{T'} \sum_{t=1}^{T'} 1[x^{(t)} \leq x^*] \geq \delta$ where $x^{(t)}$ is the user state in the t^{th} interaction.

Definition C.2. Given a simple policy π, and some $\delta > 0$, we say that user state *lies in range* $\mathbf{x} \subseteq \mathbb{R}$ *for a δ-fraction of time* if there exists a constant T such that for all $T' \geq T$, $\frac{1}{T'} \sum_{t=1}^{T'} 1[x^{(t)} \in \mathbf{x}] \geq \delta$.

The following is a more general statement of Theorem 4.5.

Theorem C.3. *If an app sees in its user an achievable payoff of at least*

$$\max_{\pi} J(\pi) \in \Omega\left(\min_{\varepsilon > 0}\left\{\frac{\varepsilon \cdot \log(1/\gamma)}{\delta(\log \widetilde{f}(x+\varepsilon) - \log \widetilde{f}(x))} + \frac{1}{1 - \widetilde{f}(x)^{\delta/2} \cdot \gamma}\right\}\right),$$

any app policy where user state is (x, δ) bounded is suboptimal. Here, Ω treats content attributes (i.e., $\{E_i\}_{i \in \mathcal{I}}, \{R_i\}_{i \in \mathcal{I}}$) as constants.

Due to the constants in the Ω, this sufficient condition is non-trivial when there is a user state x^* that results in very low user demand and is a small constant distance below the user state x. Please see (6) for explicit constants.

Proof of Theorem C.3. In this proof, we consider all policies that result in users not having a user state of at least x a constant δ fraction of the time and examine whether they can be optimal. Our proof is roughly as follows. We can immediately rule out policies that drive user state to such a low level that—even if the app maximizes revenue at every interaction—user demand is so low that the policies must be suboptimal (Lemma C.4). All remaining policies must keep user state within a bounded range for a constant fraction of the time. Suppose to the contrary that one of these policies π is optimal and that the optimal payoff—which π must realize—is sufficiently large. Then the large payoff of π can be further increased by first raising user state to a higher level (Lemma C.6), contradicting the optimality of π.

Lemma C.4. *Suppose that for some finite level of user state x and $\delta > 0$*

$$\max_{\pi} J(\pi) \in \Omega\left(\frac{\max_{i\in\mathcal{I}} \mathbb{E}\left[R_i\right]}{1-\widetilde{f}(x)^{\delta}\cdot\gamma}\right).$$

Then, any policy where user state is less than x a δ-fraction of the time is suboptimal.

Proof. Let $K \coloneqq \max_{i\in\mathcal{I}} \mathbb{E}\left[R_i\right]$, which exists by compactness of $\{\mathbb{E}\left[R_i\right]\}_{i\in\mathcal{I}}$. Fix a policy where users are less than x^* satisfied for at least a δ-fraction of the time. By definition, there exists a constant T such that for all $T' \geq T$, $\mathrm{Avg}(\left\{1[x^{(t)} \leq x^*]\right\}_{t\leq T'}) \geq \delta$. We can thus write $J(\pi)$ as

$$\begin{aligned}
J(\pi) &= \sum_{t=1}^{T-1}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)r_t + \sum_{t=T}^{\infty}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)r_t \\
&\leq \sum_{t=1}^{T-1}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)K + \sum_{t=T}^{\infty}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)K.
\end{aligned}$$

Let $S_t = \left\{\tau \in [2,t] : x^{(\tau)} \leq x\right\}$ denote the subset of the first t timesteps where the user state is no more than x. Since $\widetilde{f} \leq 1$ and $\widetilde{f}$ is monotonically non-decreasing, we can further simplify

$$\begin{aligned}
J(\pi) &\leq \sum_{t=1}^{T-1}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)K + \sum_{t=T}^{\infty}\left(\prod_{\tau\in S_t}\widetilde{f}(x^*)\right)\gamma^{t-1}K && (\widetilde{f}\text{ is monotonically non-decreasing}) \\
&= \sum_{t=1}^{T-1}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)K + \sum_{t=T}^{\infty}\widetilde{f}(x^*)^{|S_t|}\gamma^{t-1}K && (\delta\text{-fraction user state assumption}) \\
&\leq \sum_{t=1}^{T-1}\left(\prod_{\tau\in[2,t]}\widetilde{f}(x^{(\tau)})\gamma\right)K + \sum_{t=T}^{\infty}\widetilde{f}(x^*)^{(\delta+o(1))(t-1)}\gamma^{t-1}K && \left(\frac{1}{t}\sum_{\tau=1}^{t}1[x^{(\tau)} \leq x^*] \geq \delta + o(1)\right) \\
&\leq O\left(TK\right) + \sum_{t=1}^{\infty}\widetilde{f}(x^*)^{\delta(t-1)}\gamma^{t-1}K.
\end{aligned}$$

Thus, if

$$\max_{\pi} J(\pi) \in \Omega\left(TK + \frac{K}{1-\widetilde{f}(x)^{\delta}\cdot\gamma}\right) = \Omega\left(\frac{K}{1-\widetilde{f}(x)^{\delta}\cdot\gamma}\right),$$

any policy where users are less than x^* satisfied a δ of the time cannot be optimal. □

Before proceeding to the next step of the proof, we first characterize the payoff of modifying a policy to first pre-emptively increase user state.

Lemma C.5. *Given a policy π, let π' denote the policy where the app repeatedly shows a content $i^* \in \mathcal{I}$ where $\mathbb{E}\left[E_{i^*}\right] > 0$ until reaching some fixed user state x_{target} at which it switches to running the policy π. Letting C_1, C_2 denote positive constants, the payoff of π' is*

$$J(\pi') \geq (C_1\gamma)^{\frac{x_{target}}{\mathbb{E}[E_{i^*}]}}\cdot V_{\pi}(x_{target}) - C_2\frac{x_{target}}{\mathbb{E}\left[E_{i^*}\right]}. \tag{5}$$

Proof. Given a transcript of app-user interactions H, let $T_H \coloneqq \min_{t\in\mathbb{N}} \left\{x^{(t)} + e^{(t)} \geq x_{\text{target}}\right\}$ denote the timestep at which user state first passes the level of x_{target}. We then define a function g that edits transcripts by setting $g(H) = \emptyset$ if $T = \emptyset$ and $g(H) = \left\{(1, 0, \sum_{t=1}^{T-1} e^{(t)}, \emptyset)\right\} + \left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_{t\geq T}$ otherwise. We then define the policy π' as $\pi'(H) = i^*$ if $H = \emptyset$ otherwise $\pi'(H) = \pi(g(H))$.

We can decompose the payoff of the policy π' into

$$\begin{aligned}
J(\pi') &= \sum_{t=1}^{T-1} \mathbb{E}\left[\left(\prod_{\tau=2}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right] + \mathbb{E}\left[\left(\prod_{\tau=2}^{T} \gamma\widetilde{f}(x^{(\tau)})\right) \sum_{t=T}^{\infty} \left(\prod_{\tau=T}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right] \\
&= \underbrace{\sum_{t=1}^{T-1} \mathbb{E}\left[\left(\prod_{\tau=2}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right]}_{A} + \underbrace{\mathbb{E}\left[\left(\prod_{\tau=2}^{T} \gamma\widetilde{f}(x^{(\tau)})\right)\right]}_{B} \mathbb{E}\left[\sum_{t=T}^{\infty} \left(\prod_{\tau=T}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right] \\
&= \sum_{t=1}^{T-1} \mathbb{E}\left[\left(\prod_{\tau=2}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right] + \mathbb{E}\left[\left(\prod_{\tau=2}^{T} \gamma\widetilde{f}(x^{(\tau)})\right)\right] V_\pi(x^{(T)}) \\
&\geq \sum_{t=1}^{T-1} \mathbb{E}\left[\left(\prod_{\tau=2}^{t} \gamma\widetilde{f}(x^{(\tau)})\right) r_t\right] + \mathbb{E}\left[\left(\prod_{\tau=2}^{T} \gamma\widetilde{f}(x^{(\tau)})\right)\right] V_\pi(x_{\text{target}}),
\end{aligned}$$

where the inequality follows because the value function is monotonically non-decreasing. We can interpret A as the cost incurred from showing content i^* in the initial demand-building phase and B as the discount penalty incurred from spending time building user demand.

We first observe that T—the timestep at which user state reaches the level of x_{target}—corresponds to the stopping time of a random walk. In particular, we know that T is a stopping time for the filtration $\{\mathcal{F}_t\}$ generated by the realized user experiences $\left\{e^{(t)}\right\}_{t\in\mathbb{N}}$. We also know that the difference sequence

$$Y_t \coloneqq (x^{(t)} - x^{(0)}) - t \cdot \mathbb{E}\left[E_{i^*}\right]$$

is a martingale with respect to $\mathcal{F}_t$ from $t = 1, \ldots, T$. Moreover, since the variance of user experiences are bounded by assumption, we know that

$$\mathbb{E}\left[\left|x^{(t)} - x^{(t-1)} - \mathbb{E}\left[E_{i^*}\right]\right|\right] = \mathbb{E}\left[\left|e^{(t)} - \mathbb{E}\left[E_{i^*}\right]\right|\right] < \infty.$$

If we cap the stopping time T by some constant $n \in \mathbb{Z}$, which we will denote by $T \wedge n \coloneqq \min\{T, n\}$, then we can also observe that $T \wedge n$ is also a stopping time for $\mathcal{F}$. Since n is finite and thus $\mathbb{E}\left[T \wedge n\right] < \infty$, we appeal to the optional stopping theorem to observe that

$$x^{(0)} = \mathbb{E}\left[x^{(T\wedge n)}\right] - \mathbb{E}\left[(T \wedge n) \cdot E_{i^*}\right].$$

We can upper bound how much we are expected to overshoot the target user state x_{target} by

$$\mathbb{E}\left[x^{(T\wedge n)}\right] \leq x_{\text{target}},$$

which directly implies that

$$\mathbb{E}\left[T \wedge n\right] \leq x_{\text{target}} / \mathbb{E}\left[E_{i^*}\right].$$

The monotone convergence theorem then implies that the expected stopping time is $\mathbb{E}\left[T\right] = \frac{x_{\text{target}}}{\mathbb{E}[E_{i^*}]}$.

We now lower bound the A summand, which corresponds to the revenue realized prior to timestep T. If the content i^* does not result in negative revenue for the app creator, i.e., $\mathbb{E}\left[R_{i^*}\right] \geq 0$, we can

bound $A \geq 0$. Otherwise, if the content does cause negative revenue, i.e. $\mathbb{E}[R_{i^*}] < 0$, then we can lower bound $A \geq \mathbb{E}[T]\,\mathbb{E}[R_{i^*}]$. Thus, we can lower bound, $A \geq \frac{x_{\text{target}}}{\mathbb{E}[E_{i^*}]} \min\{0, \mathbb{E}[R_{i^*}]\}$, meaning the constant C_2 in the lemma statement can therefore be understood as $C_2 = \min\{0, \mathbb{E}[R_{i^*}]\}$.

To bound the term B, we note that the probability of user engagement is bounded away from zero by some constant. Formally, there exists some $C > 0$ such that for all $x \in \mathbb{R}$, $f(x) \geq C$ and thus also a $C_1 > 0$ such that for all $x \in \mathbb{R}$, $\widetilde{f}(x) \geq C_1$. We can thus bound $\prod_{\tau=2}^{T} \gamma \widetilde{f}(x^{(\tau)}) \geq (\gamma C_1)^T$. By Jensen's inequality,

$$B = \mathbb{E}\left[(\gamma C_1)^T\right] \geq (\gamma C_1)^{\mathbb{E}[T]} \geq (\gamma C_1)^{\frac{x_{\text{target}}}{\mathbb{E}[E_{i^*}]}}.$$

That $J(\pi) \geq 0$ concludes our proof. □

We can always improve policies that keep user state within a small range and attain a high payoff.

Lemma C.6. *Consider any closed range of user states* $\mathbf{x}$ *(where* $\underline{\mathbf{x}} = \min \mathbf{x}$ *and* $\overline{\mathbf{x}} = \max \mathbf{x}$*) and any policy* π *with* $J(\pi) > 0$*, where user state lies within* $\mathbf{x}$ *for a* $\delta > 0$ *fraction of the time. The policy* π *is suboptimal if*

$$J(\pi) \in \Omega\left(\min_{x_{imp} \geq \max \mathbf{x}} (x_{imp} - \underline{\mathbf{x}}) \left(C + \frac{\log(1/\gamma)}{\delta \log(\widetilde{f}(x_{imp})/\widetilde{f}(\overline{\mathbf{x}}))}\right)\right).$$

Proof. We will improve upon policy π by defining a new policy π'. Fix any $x_{\text{imp}} \geq \max \mathbf{x}$. By assumption, there exists an content i^* with positive user effect: $\mathbb{E}[E_{i^*}] > 0$. The policy will show this content i^* until user state is raised to a target level of user state of $x_{\text{target}} = x_{\text{imp}} - \underline{\mathbf{x}}$. We will then execute the original policy π as if we had never raised user state to x_{target}.

Formally, given a transcript H, let $T_H \coloneqq \min_{t \in \mathbb{N}} \left\{x^{(t)} + e^{(t)} \geq x_{\text{target}}\right\}$ denote the timestep at which user state first passes the level of x_{target}. We then define the transcript editing function g where $g(H) = \emptyset$ if $T = \emptyset$ and otherwise $g(H) = \{(1, 0, 0, \emptyset)\} + \left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_{t \geq T}$. We then define the policy π' as $\pi'(H) = i^*$ if $H = \emptyset$ otherwise $\pi'(H) = \pi(g(H))$. For the remainder of this proof, we use $\widehat{x}_{\text{target}} = x^{(T)}$ to denote the user state that policy π' achieves before switching to simulating π; note that $\widehat{x}_{\text{target}} \geq x_{\text{target}}$. By Lemma C.5, we can lower bound the payoff of π' by

$$\underbrace{C_1 \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}}_{A} + \underbrace{(K\gamma)^{\frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}}}_{B} \underbrace{\mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t \sim \pi} \left[\sum_{t=2}^{\infty} \left(\prod_{\tau=2}^{t} \widetilde{f}(x^{(\tau)} + x_{\text{target}})\gamma\right) r_t\right]}_{C},$$

where $K > 0$ is a constant that lower bounds the range of $\widetilde{f}$. We now turn to lower bounding C. We introduce a sequence of constants $\left\{C^{(\tau)}\right\}_{\tau \in \mathbb{N}}$, where we guarantee that $C^{(\tau)} \geq 1$ for all $\tau \in \mathbb{N}$. We can observe that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t \sim \pi} \left[\sum_{t=1}^{\infty} \left(\prod_{\tau=2}^{t} C^{(\tau)} \widetilde{f}(x^{(\tau)})\gamma\right) r_t\right] \\
&= \mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t \sim \pi} \left[\sum_{t=1}^{k-1} \left(\prod_{\tau=2}^{t} C^{(\tau)} \widetilde{f}(x^{(\tau)})\gamma\right) r_t + \left(\prod_{\tau=2}^{k-1} C^{(\tau)} \widetilde{f}(x^{(\tau)})\gamma\right) C_k J_{x^{(k)}}(\pi)\right]
\end{aligned}$$

is non-decreasing in C_k. Now, we let $C^{(\tau)} = \frac{\widetilde{f}(x^{(\tau)} + \widehat{x}_{\text{target}})}{\widetilde{f}(x^{(\tau)})}$ and will resort to two lower bounds. Since f is monotonically non-decreasing, we immediately know that $C^{(\tau)} \geq 1$ for all $\tau \in \mathbb{N}$. Using the

shorthand $\eta = \widetilde{f}(x_{\text{imp}})/\widetilde{f}(\overline{\mathbf{x}})$ and recalling that $\widetilde{f}$ is monotonically non-decreasing, we will also use the following lower bound for every $x \in \mathbf{x}$,

$$\begin{aligned}
\widetilde{f}(x + x^{(T-1)} + e^{(T-1)}) &\geq \widetilde{f}(x^{(\tau)} + x_{\text{target}}) && (x^{(T-1)} + e^{(T-1)} \geq x_{\text{imp}}) \\
&\geq \widetilde{f}(x_{\text{imp}}) && (x \geq \underline{\mathbf{x}}) \\
&\geq \eta \widetilde{f}(x). && (\overline{\mathbf{x}} \geq x)
\end{aligned}$$

Recall that, by assumption, for every $\xi > 0$, there is a timestep T' where all $k \geq T'$ satisfy $\frac{1}{k}\sum_{t=1}^{k} 1[x^{(t)} \in \mathbf{x}] \geq \delta$. Putting everything together, we conclude that for all $k \geq T'$,

$$\begin{aligned}
\prod_{\tau=1}^{k} C^{(\tau)} &= \left(\prod_{\tau=1}^{k} (C^{(\tau)})^{1[C^{(\tau)} \in \mathbf{x}]}\right)\left(\prod_{\tau=1}^{k} (C^{(\tau)})^{1[C^{(\tau)} \notin \mathbf{x}]}\right) \\
&\geq \left(\prod_{\tau=1}^{k} \eta^{1[C^{(\tau)} \in \mathbf{x}]}\right)\left(\prod_{\tau=1}^{k} 1^{1[C^{(\tau)} \notin \mathbf{x}]}\right) \geq \eta^{k\delta}.
\end{aligned}$$

We thus have that for any $k \geq T'$, taking an expectation over the trajectory of policy π,

$$\begin{aligned}
C &= \mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t}\left[\sum_{t=1}^{\infty}\left(\prod_{\tau=2}^{t} C^{(\tau)} \widetilde{f}(x^{(\tau)})\gamma\right) r_t\right] \\
&\geq \mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t}\left[\sum_{t=1}^{k-1}\left(\prod_{\tau=2}^{t} \widetilde{f}(x^{(\tau)})\gamma\right) r_t\right] + \eta^{\delta k} \mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t}\left[\sum_{t=k}^{\infty}\left(\prod_{\tau=2}^{t} \widetilde{f}(x^{(\tau)})\gamma\right) r_t\right] \\
&\geq \eta^{\delta k} \mathop{\mathbb{E}}_{\left\{(s^{(t)}, r^{(t)}, e^{(t)}, i^{(t)})\right\}_t}\left[\sum_{t=1}^{\infty}\left(\prod_{\tau=2}^{t} \widetilde{f}(x^{(\tau)})\gamma\right) r_t\right] - k \max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right] \\
&\geq \eta^{\delta k} J(\pi) - k \max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right].
\end{aligned}$$

Simplifying our initial equality, we have

$$\begin{aligned}
J(\pi') - J(\pi) &\geq C_1 \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}\left[E_{i^*}\right]} + \left((C_2\gamma)^{\frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}} \eta^{\delta k} - 1\right) J(\pi) - (K\gamma)^{\left(1 + \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}\right)} k \max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right] \\
&\geq C_1 \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}\left[E_{i^*}\right]} + \left((K\gamma)^{\frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}} \eta^{\delta k} - 1\right) J(\pi) - k \max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right]
\end{aligned}$$

where the second inequality uses the fact that $B \leq 1$. Choosing

$$k \coloneqq \frac{1}{\delta \log(\eta)} \log\left(2\left(K\gamma\right)^{-\frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}[E_{i^*}]}}\right) + T' = \frac{1}{\delta} \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}\left[E_{i^*}\right]} \log(1/K\gamma) + \frac{1}{\delta}\log(2) + T',$$

we have for some appropriate constant C

$$\begin{aligned}
J(\pi') - J(\pi) &\geq C_1 \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}\left[E_{i^*}\right]} + J(\pi) \\
&\quad + \left(\frac{1}{\delta \log(1/\eta)} \frac{x_{\text{imp}} - \underline{\mathbf{x}}}{\mathbb{E}\left[E_{i^*}\right]} \log(1/K\gamma) + \frac{1}{\delta}\log(2) + T'\right) \max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right] \\
&= J(\pi) + C(x_{\text{imp}} - \underline{\mathbf{x}})\left(C + \frac{\log(1/K\gamma)}{\delta \log(\eta)}\right).
\end{aligned}$$

□

The following fact says that we can combine the results of Lemma C.4 and Lemma C.5 to assert our claim.

Fact C.7. *Every policy that is less than x satisfied a δ fraction of the time is either within $[x', x]$ satisfied a $\delta/2$ fraction of the time or less than x' satisfied a $\delta/2$ fraction of the time.*

Proof. Fix any two choices of user states x and x' where $x > x'$ and $\delta > 0$. Given a policy and $k \in \mathbb{N}$, let $T_{1,k}$ denote the fraction of the first k timesteps where user state is at most x, let $T_{2,k}$ denote the fraction of the first k timesteps where user state is less than x', and let $T_{3,k}$ denote the fraction of the first k timesteps where user state is within $[x', x]$. We have that for all $k \in \mathbb{N}$, $T_{1,k} = T_{2,k} + T_{3,k}$. Now, consider the set S of all policies where user state is less than x a δ fraction of the time and the set S' of policies where user state lies in $[x', x]$ a $\delta/2$ fraction of the time. That is, where after some constant T', for all $k \geq T'$, $T_{1,k} \geq \delta$ and $T_{3,k} \leq \delta/2$ and thus $T_{2,k} \geq \delta/2$. □

We thus have that, for any choice of δ and $x > x'$, if

$$J(\pi) \in \Omega\left(\min_{x_{\text{imp}} \geq x} (C + x_{\text{imp}} - x') \left(\frac{2\log(1/K\gamma)}{\delta \log(\widetilde{f}(x_{\text{imp}})/\widetilde{f}(x))}\right) + \frac{\max_{i \in \mathcal{I}} \mathbb{E}\left[R_i\right]}{1 - \widetilde{f}(x')^{\delta/2} \cdot \gamma}\right), \tag{6}$$

then all policies where users are less than x satisfied a δ fraction of the time are suboptimal. □

C.2 Proof of Proposition 4.2

Proposition 4.2. In Example 4.1, for an appropriate choice of a, b, the user is strictly less satisfied by the optimal app policy when there is less friction compared to when there is more friction. Formally, for any $c' > c$, let x_t^c and $x_t^{c'}$ denote the user experience states at time step t in the optimal policies for friction parameters c and c'. Then, for all t, $x_t^c \leq x_t^{c'}$ and this inequality is strict for $t \geq 2$.

Proof. We will write $\widetilde{f}$ with subscripts $\widetilde{f}_c$ and $\widetilde{f}_{c'}$ to denote the operative friction constant. Let us fix $a = 0$. We can write our user's demand function as $f(x) = 0.6 \cdot 1[0 \leq x < b] + 0.99 \cdot 1[x \geq b]$. It has two discontinuities: one at $x = 0$ and one at $x = 4$. Both are reachable with a single interaction by showing content $i = 0$ and $i = b$ respectively. We now set b as follows. Consider the function

$$\begin{aligned} g(c^*) &= \tfrac{1}{1-\gamma\widetilde{f}_{c^*}(b)} - \tfrac{1}{1-\gamma\widetilde{f}_{c^*}(0)} \\ &= \frac{1}{1 - 0.9(0.99 + \frac{1-0.99}{1-0.9(1-(1-c^*)0.99)}(1-c^*) \cdot 0.99 \cdot 0.9)} \\ &\quad - \frac{1}{1 - 0.9(0.6 + \frac{1-0.6}{1-0.9(1-(1-c^*)0.6)}(1-c^*) \cdot 0.6 \cdot 0.9)}. \end{aligned}$$

g is positive and monotonically increasing for $c^* \in [0, 1]$. We can therefore let

$$\varepsilon \coloneqq \min\left\{\tfrac{1}{2}(g(c') - g(c)), 0.1\right\} > 0$$

and $b = g(c) + \varepsilon$. By Lemma 3.1, we know that the optimal app policy must be one of three:

1. Maximally decrease user state at each interaction by repeatedly showing content $i_t = b$.
2. Show content $i_t = 0$ in perpetuity to maintain the user state at $x = 0$.
3. Show content $i_1 = -b$ and subsequently show content $i_t = 0$ in perpetuity to maintain the user state at $x = b$.

The payoff of the first policy is simply $1 + b$, as the user will stop interacting immediately.

Now suppose the friction constant is c. By Theorem 2.2, the payoff of the second policy is $\frac{1}{1-\gamma\widetilde{f}_c(0)}$ while the payoff of the third policy is $\frac{1}{1-\gamma\widetilde{f}_c(b)} - b = \frac{1}{1-\gamma\widetilde{f}_c(0)} - \varepsilon$. This means the app will always prefer the second policy over the third. If instead the friction constant is c', by Theorem 2.2, the payoff of the second policy is $\frac{1}{1-\gamma\widetilde{f}_{c'}(0)}$ while the payoff of the third policy is

$$\frac{1}{1-\gamma\widetilde{f}_{c'}(b)} - b = \frac{1}{1-\gamma\widetilde{f}_{c'}(0)} + g(c') - g(c) - \varepsilon > \frac{1}{1-\gamma\widetilde{f}_{c'}(0)},$$

where we use that $g(c) < g(c')$. In this case, we have that the app prefers the third policy over the second.

Since the utility of the first policy is friction-independent and the other two policy utilities are decreasing in friction, e can observe that the first policy is always suboptimal as

$$\frac{1}{1-\gamma\widetilde{f}_{c'}(b)} - g(c) \geq 1 + g(c) + 0.5$$

for all $c \leq 0.3$ and so

$$\frac{1}{1-\gamma\widetilde{f}_{c'}(b)} - b \geq 1 + b.$$

Thus, under friction c', the app's optimal policy is policy 3, resulting in the user state trajectory $x_t = b$ for all $t \geq 2$. Similarly, under friction c, the app's optimal policy is policy 2, resulting in the user state trajectory $x_t = 0$ for all $t \geq 2$. It follows that $x_t^c < x_t^{c'}$ for all $t \geq 2$. □

C.3 Analog of Proposition 4.2 for Absent/Complete Friction

The app's increased investment in user engagement makes it possible for usage of an app to increase when friction is increased. If the user gains utility from its interactions with the app, the user's discounted utility may also be higher. This holds true even in the extreme case where we are comparing the absence of friction ($c = 0$) with the complete friction ($c = 1$) of a user being entirely banned from rejoining apps.

Proposition C.8. In Example 4.1, the number of expected app-user interactions in the first $N = 100$ timesteps is lower when there is less friction than when there is more friction. Similarly, if the user gains utility upon every app interaction, its cumulative utility (for discount factor $\gamma' \in (0, 0.98)$) is strictly greater when there is more friction.

By replacing Example 4.1's demand levels of 60% and 99% with appropriate substitutes, Proposition C.8 can hold for arbitrary choices of $N \in \mathbb{Z}$ and $\gamma' \in (0, 1)$. We can also extend Example 4.1 to generalizations of our basic model of user-app interactions. For example, the preceding Proposition 4.2 and Proposition C.8 both hold if instead of the user's satisfaction being the sum of previous experiences, the user's satisfaction corresponds to the average or arbitrarily discounted sum of their previous experiences.

Proof of Proposition C.8. Let us resume our construction from the proof of Proposition 4.2. In this construction, when there is less friction, the app's policy results in a user state trajectory such that $f(x_t) = 0.6$ for all $t \geq 2$. When there is more friction, the app's policy results in $f(x_t) = 0.99$ for all $t \geq 2$.

Let us first upper bound the expected number of interactions that occur in the first $N = 100$ timesteps under less friction. For this upper bound, we can assume a frictionless setting, in which

case the expected number of times the user engages the app is $1 + 0.6(N-1) = 60.4$; here, the plus-1 reflects the guaranteed interaction at the first timestep. To lower bound the expected number of interactions that occur under more friction, we can consider the opposite extreme and suppose complete friction: in this case, the expected number of interactions is $\sum_{i=1}^{N} b^{n-1} = \frac{1-0.99^N}{1-0.99} = 63.4$. Thus, there is always a greater expected number of interactions in the first $N = 100$ timesteps when there is more friction.

Now suppose the user receives a reward of $R > 0$ if it interacts with the app and no reward otherwise. To upper bound the user's utility when there is less friction, assume a frictionless setting and observe that the user's utility is then $\sum_{t=1}^{\infty}(\gamma')^{t-1}(0.6R) = \frac{0.6R}{1-\gamma'} = 30$. To lower bound the user's utility when there is more friction, assume a full friction setting and observe the user's utility is $\sum_{t=1}^{\infty}(\gamma' 0.99)^{n-1}R = \frac{R}{1-\gamma' 0.99} = 33.5$. Thus, the user's utility is also higher under more friction. □

C.4 A Second Example of Friction

Example C.9. Suppose a user is repeatedly choosing an app to use from an ecosystem of competitors, which we will treat as a mean-field. We zoom in on the point-of-view of a specific app, which we will say is Instagram.

Original scenario: Suppose the user's interest in Instagram can be represented as a numerical score that falls into one of three levels:

- The user is entirely disinterested if their interest in Instagram is below a threshold of 0. In this case, the user will stop interacting with Instagram.
- The user exercises healthy usage of Instagram if their interest in Instagram falls within the interval $[0, 6)$. In this case, the user has a 60% chance of using Instagram on any given day.
- The user is addicted to Instagram if their interest in Instagram is larger 6. In this case, the user has a 90% chance of using Instagram on any given day.

Further suppose Instagram creator has a $\gamma = 0.95$ discount factor and a linear content landscape parameterized by $i \in [-6, 6]$ such that displaying content i yields $R_i = 1 + i$ revenue for Instagram and a $E_i = -i$ effect on user interest.

Alternate scenario: Let us consider again the original scenario, but imagine Instagram's competitors begin serving increasingly addictive content. To reflect that Instagram's users might get addicted to a competitor while they're not using Instagram, if a user does not open Instagram for a day, the probability they use it the next day is reduced by 50%. Let us also suppose the competitors' new use of addictive content has a side-effect of disinclining users from starting new sessions with them, increasing the retention rate on Instagram by some w% (can set $w = 0$). As a result, the competition has gotten stronger at increasing session length but also weaker at attracting new sessions. The engagement probabilities in this alternate scenario can thus be summarized as

- Interest below 0: users interact with 0% chance.
- Interest in $[0, 6)$: The user has a 60% + w% chance of staying on Instagram if they are already on it, but a 30% chance of using Instagram if they are not already.
- Interest above 6: The user has a 90% + w% chance of staying on Instagram if they are already on it, but a 45% chance of using Instagram if they are not already.

If we compare Instagram's optimal policy and the resulting engagement frequency, we see that the alternate scenario (where competitors become more addictive) changes the optimal policy of Instagram to be more addictive and ends up increasing the number of user-app interactions.

Proposition C.10. In Example C.9, for all $w \in [0\%, 5\%]$, the user has a higher interest in Instagram in the alternate scenario when competitors are addictive. Formally, let x_t and x'_t denote the user's interest in Instagram at time step t in the original and alternate scenario respectively. Then, for all t, $x_t \leq x'_t$ and this inequality is strict for $t \geq 2$. Moreover, for any time period, the expected number of days that the user spends on Instagram is strictly higher in the alternate scenario.

Proof. First, we can observe that both scenarios correspond to an instance of our model where the app creator's content decision problem is linear and the user demand function has two discontinuities. By Lemma 3.1, we know the optimal app policy in either scenario must be one of three possibilities:

1. π_1: Maximally decrease user interest by showing content $i^{(1)} = 6$. No interactions occur after.
2. π_2: Show content $i^{(t)} = 0$ at all timesteps t, keeping interest at $x^{(t)} = 0$.
3. π_3: First show content $i^{(1)} = -6$ then show $i^{(t)} = 0$ thereafter, keeping interest at $x^{(t)} = 6$.

The first policy only obtains revenue in the first timestep, giving a payoff of 7 in both scenarios: $J(\pi_1) = J'(\pi_1) = 7$, where we use J' to denote the payoff in the alternate scenario. We can manually compute the payoff of the remaining two policies using Theorem 2.2.

In the original scenario, the payoffs of the second policy and third policy are, respectively,

$$J(\pi_2) = \sum_{t=1}^{\infty} \left(\gamma \cdot 0.6 + \gamma^2 \frac{1 - 0.6}{1 - \gamma(1 - 0.6)} \cdot 0.6\right)^{t-1} = 12.4,$$
$$J(\pi_3) = -6 + \sum_{t=1}^{\infty} \left(\gamma \cdot 0.9 + \gamma^2 \frac{1 - 0.9}{1 - \gamma(1 - 0.9)} \cdot 0.9\right)^{t-1} = 12.1,$$

meaning that the second policy is optimal. Thus, the user states—and equivalently user interest in Instagram—are $x_t = 0$ for all $t \geq 2$ under the original scenario.

In the alternate scenario, the payoffs of the second policy and third policy are, respectively,

$$J'(\pi_2) = \sum_{t=1}^{\infty} \left(\gamma(0.6 + w) + \gamma^2 \frac{1 - (0.6 + w)}{1 - \gamma(1 - 0.3)} \cdot 0.3\right)^{t-1} \approx \frac{1}{0.1067 - 0.1418w},$$
$$J'(\pi_3) = -6 + \sum_{t=1}^{\infty} \left(\gamma(0.9 + w) + \gamma^2 \frac{1 - (0.9 + w)}{1 - \gamma(1 - 0.45)} \cdot 0.45\right)^{t-1} \approx \frac{1}{0.05995 - 0.09948w} - 6.$$

Since $\frac{1}{0.05995 - 0.09948w} - 6 > \frac{1}{0.1067 - 0.1418w} > 7$ for all $w \in [0, 0.05]$, the third policy is optimal. Thus, the user states—and equivalently user interest in Instagram—are $x_t = 6$ for all $t \geq 2$ under the original scenario.

We now turn to the second claim. Fix any timestep $t \geq 2$. In Scenario 1, the probability that the user interacts with Instagram is 60%, i.e. $\Pr(s_t = 1) = f(0) = 0.6$. In Scenario 2, we will lower bound the probability of interaction by 80%, i.e. $\Pr(s_t = 1) > 0.8$. To show this, suppose to the contrary that there is a timestep $t' \geq 2$ where $\Pr(s_{t'} = 1) \leq 0.8$. Letting t' be the smallest such timestep, the observation that $\Pr(s_{t'-1} = 1) > 0.8$ directly leads to a contradiction:

$$\Pr(s_{t'} = 1) = (0.9 + w) \cdot \Pr(s_{t'-1} = 1) + 0.45 \cdot \Pr(s_{t'-1} = 0) \geq 0.9 \cdot 0.8 + 0.45 \cdot 0.2 = 0.81,$$

confirming $\Pr(s_t = 1) > 0.8$ for all $t \geq 2$. The second claim then follows by linearity of expectation. □

As an illustration of what this example demonstrates, suppose Instagram and a competitor app both implement a recommendation algorithm that prioritizes interesting but not addictive content. If the competitor app suddenly improves the quality of its recommendation algorithm, standard models of competition tell us that Instagram may need to show fewer ads to become more competitive for user engagement.

However, suppose instead that the competitor app does not improve its algorithm. Rather, it switches to recommending extremely addictive content that results in longer user sessions but also repels users, resulting in fewer new user sessions. Our model provides a formal treatment of this scenario, which is not as easily captured by standard models of competition. Proposition C.10 says that even though the competitor has become less effective at attracting users, by becoming more addictive, the competitor might still incentivize Instagram to show fewer ads to compete for user engagement. In fact, Proposition C.10 says something stronger: Instagram may end up sacrificing so much of its profit to increase user engagement that it sees more usage than it did prior.