Social Learning from Online Reviews with Product Choice

Costis Maglaras
Columbia Business School, Columbia University, New York, USA, c.maglaras@gsb.columbia.edu
Marco Scarsini
Dipartimento di Economia e Finanza, Luiss University, Rome, Italy, marco.scarsini@luiss.it
Dongwook Shin
The HKUST Business School, Clear Water Bay, Kowloon, Hong Kong, dwshin@ust.edu.hk
Stefano Vaccari
Global Data Hub, Global Digital Solutions, ENEL Global Services s.r.l., Rome, Italy, stefano.vaccari@enel.com

This paper studies product ranking mechanisms of a monopolistic online platform in the presence of social learning. The products' quality is initially unknown, but consumers can sequentially learn it as online reviews accumulate. A salient aspect of our problem is that consumers, who want to purchase a product from a list of items displayed by the platform, incur a search cost while scrolling down the list. In this setting, the social learning dynamics, and hence the demand, is affected by the interplay of two unique features: substitution and ranking effects. The platform can influence the social learning dynamics by adjusting the ranking of the products to ultimately maximize the revenue collected from commission fees for sold items. To formulate the problem in a tractable form, we use a large-market (fluid) approximation and show that consumers eventually learn the products' quality and characterize the speed of learning. Armed with this backing, we formulate the platform's ranking problem in the fluid setting, where we assume the perspective of an uninformed platform that does not know the true quality vector but rather learns it through consumers' review process. We compare different ranking policies based on the worst-case regret with respect to a fully-informed platform benchmark. Our analysis yields three main insights. First, a greedy policy that maximizes immediate revenue by displaying products based on current ratings may incur highly suboptimal worst-case regret, as it may relegate the most profitable products to the lowest positions in the ranking if their current rating is not high enough. Second, a simple variant of the greedy policy can sufficiently alleviate the regret by balancing the trade-off between exploration and exploitation. Third, we characterize the critical level of search cost for which the regret does not grow exponentially with the number of products.

Key words: social learning, information aggregation, online reviews, online platforms

1. Introduction

1.1. Motivation

It is common for consumers to rely on online review platforms, such as Amazon, TripAdvisor, Yelp, IMDb, etc., when planning to purchase a new product (or service). In such online environments, owing to a marked increase in product variety and complexity, consumers often refer to online reviews

to estimate the quality of a product. As consumers observe and report online reviews over time, they engage in what is called a social learning process; namely, through online reviews, consumers share their experiences and opinions about products with other consumers. As more reviews accumulate, consumers can have a more reliable estimate of the quality of different products to make better-informed purchase decisions. In the past decades, online reviews have gained an increasingly central role in the consumer decision process. As a consequence, the literature in operations management and economics has recently dealt with review-based social learning problems. However, despite the fact that users typically choose among many competing alternatives in online marketplaces, most academic studies of social learning concern single-product settings. In this paper, we shed light on new challenges that platforms face when social learning involves consumers' choice among multiple products. In particular, we focus on the challenges that emerge from the interplay of two unique features: *substitution effects* and *ranking effects* among displayed products.

In the presence of product choice, the social learning processes are correlated across products: information accumulates faster for products that consumers perceive as more appealing in terms of online ratings and price, as these products will be selected more frequently and produce more reviews. In other words, alongside the usual questions on what the final outcomes of the social learning process are (that is, whether consumers' beliefs converge and where), when product choice is considered, it is also important to understand how these quality beliefs converge to their limiting points, that is, how learning paths interfere with each other, and how this coupling depends on the market parameters (number of alternatives, prices, true qualities, consumers' prior beliefs).

In most online review platforms, multiple items are displayed with a rank, and consumers exert more cognitive (and physical) effort while scrolling down a list of items (Kempe and Mahdian, 2008, Craswell et al., 2008, Ghose et al., 2013, Lerman and Hogg, 2014). This ranking effect can be costly for platforms: since the product's quality is typically unknown to platforms, due to possible misalignment between perceived and actual qualities, a high-quality product may be ranked low, and consumers would then require an increased cognitive effort to purchase it. Such additional effort can be modeled in the consumers' choice model as the so-called *search cost* (Stigler, 1961): due to this extra cost, high quality products may remain underexplored relative to other (possibly inferior) products. At the same time, the ranking effect gives the platform a control mechanism; that is, by picking the product ranking, the platform can affect product choice and stimulate information acquisition to ultimately maximize revenues. Without an understanding of the consumer learning dynamics and the influence of product choice, the question of how to best choose the product ranking cannot be properly addressed. In this paper, we seek to improve our understanding by modeling the aforementioned market circumstances, as we briefly describe next.

1.2. High-level Overview of the Model

We study a model of a marketplace where consumers arrive sequentially over time and decide whether to buy one of the available products or to take an outside option. Consumers are heterogeneous in their preferences towards the observable features of the products. Although initially uninformed about the intrinsic quality of the products, consumers observe the binary online reviews reported by earlier purchasers and infer the unknown quality via Bayesian updating. The probability of purchasing a product, which we refer to as the *demand function*, depends on consumers' quality estimates, their idiosyncratic preferences, and the prices of the available products; we assume Multinomial Logit (MNL) demand for our main analysis, while some of our preliminary analysis holds for a general class of demand models. Each product is characterized by a uni-dimensional quality parameter that represents the probability that a customer has a positive post-purchase experience. We assume that purchasers of a given product report "like"/"dislike" reviews that truthfully express whether they liked the product or not. Reviews are gathered and displayed to the upcoming consumers by an information aggregator, the platform.

As alluded to earlier, consumers in our model perceive an additional "search cost" which is an increasing function of the ranking in which the product is displayed by the platform and, in turn, affects the multinomial choice probabilities. Consequently, the presence of the search cost allows the platform to influence information acquisition and the learning transient by choosing the ordering in which the products are displayed to arriving consumers. In particular, the platform can boost information acquisition for a product by showing it in the top positions of the list, so that consumers can learn its quality sooner. The platform receives a constant payment from every purchase and wants to maximize cumulative revenues over a finite selling horizon. The platform is initially uninformed about the true product qualities but can learn them by dynamically adjusting the product ranking over time.

In this problem setting, the platform should simultaneously attempt to acquire information about the unknown qualities (called "exploration") and optimize the ranking decisions based on that information (called "exploitation"). This intrinsic trade-off is a salient characteristic of the ranking problem in the presence of social learning. Our goal is to illuminate two aspects in this regard: the impact of search cost on the social learning behavior of consumers; and the manner in which the ranking decisions interact with social learning. Can consumers learn the unknown quality in the presence of product choice and search cost, and if so, how fast? To what extent should the platform rely on exploration rather than on exploitation for different levels of search cost? An overarching goal of the paper is to shed light on the aforementioned questions.

1.3. Summary of the Main Results

As alluded to earlier, the interplay between the substitution and ranking effects significantly complicates the study of optimal ranking policies. To have a tractable characterization of the learning transient, we derive a large-market asymptotic (fluid) model, which provides a good approximation of the learning transient when the volume of sales is large. In this fluid model, the learning transient is described by ordinary differential equations, which are sufficiently tractable to deduce structural insights on the fundamental trade-off between exploration and exploitation in the platform's ranking problem.

We characterize the transient dynamics of the learning paths in this deterministic approximation. To isolate the substitution effect from the ranking effect, we focus on the case without search cost and provide a comparative statics analysis for how the speed of the learning transients depends on the parameters of multiple products with substitutable demand. At a high level, our findings indicate that the substitution effect is intensified by social learning: consumers tend to choose products that are more appealing in terms of review rating/price difference, which not only reduces the demand for less-appealing products but also delays their learning transients over time.

In the presence of search costs, we show that, through its ranking policy, the platform can dramatically influence the speed of learning. For example, consider a fixed (say, alphabetical) product display ordering and linearly increasing search costs. In this case, due to search costs, many low-ranked products will be very undesirable alternatives from the consumers' viewpoint. As a result, products displayed towards the end of the list will experience a slowdown as the number of products increases. In other words, as consumers take the relative positioning of products into account in their choice mechanism, consumers may need a potentially very long time to discover the high-quality products, resulting in suboptimal revenues for the platform.

For a tractable analysis of the platform's ranking problem, we continue to focus on the fluid model. In a full-information setting where the platform knows the true quality of each product, we analytically characterize the optimal ranking policy. Interestingly, the optimal ranking policy is static; that is, the optimal product ranking is fixed throughout the selling horizon, and corresponds to the ranking that maximizes the revenue rate in the full-information scenario where consumers know the products' quality. The revenue obtained by this "oracle" platform serves as an upper bound for the attainable revenue.

In our focal setting where the platform is not informed about the true quality of products, it may be beneficial for the platform to induce consumers to sufficiently explore all products (perhaps sacrificing initial revenues) in order for them to learn the products' quality faster, and to exploit this information later on. The goal of the platform is to judiciously balance the tradeoff between

exploration and exploitation to achieve the revenue that is close to the aforementioned upper bound. The performance metric we use to characterize a given ranking policy is the long-term regret, defined as the revenue gap between such policy and the previously mentioned oracle ranking policy for a sufficiently large selling horizon.

Our departure point is the *greedy* ranking policy, in which, at any point in time, products are ranked to maximize the revenue in the next period. Our analysis reveals that the greedy policy is under-explorative: in the worst case (in terms of initial quality beliefs and true quality), the regret grows with the number of products and the growth rate depends critically on the extent to which the search cost increases with product position. Our explicit characterization of the worst-case regret (Theorem 4.1) provides rough guidelines for practitioners not just in assessing the revenue loss due to the simple (and easily implementable) ranking strategy, but also in choosing the optimal number of products to display in different search environments.

Somewhat surprisingly, we show that a simple variation of the greedy policy can effectively allay the negative effect of under-exploration. This policy is referred to as *semi-greedy*: instead of assigning the top position to the most appealing product at present, the semi-greedy policy assigns the ranking based on an "index" that amalgamates the estimated quality as well as the term that represents the degree of exploration for each product. We characterize the worst-case regret under the semi-greedy policy and show, both theoretically and numerically, that the regret grows at a substantially slower rate compared to the greedy policy. By comparing the greedy and semi-greedy policies in different settings for search cost, we provide qualitative insights into the benefit of exploration relative to the extent to which the ranking effect affects consumers' decision-making.

1.4. Related Literature

Early examples of papers focusing on social learning trace back to the observational learning models studied in Banerjee (1992) and Bikhchandani et al. (1992); in a model with private signals, observable actions, and Bayesian updating, they demonstrate that rational agents may eventually ignore their private signals and decide to imitate their predecessors. Following these seminal papers, a recent stream of papers has studied social learning from consumer reviews as opposed to signals. Ifrach et al. (2019) provided sufficient conditions for perfect learning in a monopolistic market with Bayesian consumers and binary reviews, whereas Besbes and Scarsini (2018), again in a Bayesian setting, addressed the issue of self-selection bias when consumers report their ex-post utility. While the latter focused on a single-product setting, Chen et al. (2021) identified the self-selection bias in a multiproduct setting with boundedly rational consumers. Acemoglu et al. (2017) characterized the speed of learning under different rating systems in a monopolistic model of Bayesian social learning. The impact of asymmetry in learning technologies in a duopolistic market was analyzed by Kakhbod

et al. (2021), who studied a model where consumers are Bayesian and differ in the way they process the online review information.

In the field of revenue management, several papers have studied social learning from consumer reviews in various contexts; see, e.g., pricing problems (Crapis et al., 2017, Papanastasiou and Savva, 2017, He and Chen, 2017, Yang and Zhang, 2018, Shin et al., 2021, Stenzel et al., 2020); product design problems (Feldman et al., 2019, Shin and Zeevi, 2021); and information provision problem (Papanastasiou et al., 2018). In particular, our model of consumer reviews is closely related to the one in Papanastasiou et al. (2018) in that they, too, assumed that binary reviews follow a Bernoulli distribution with unknown mean and consumers use Bayesian updating (with a beta prior) to infer it. Papanastasiou et al. (2018) studied a platform's messaging policies and how they influence consumers' learning, whereas our paper concerns product ranking policies. Crapis et al. (2017) studied social learning from binary reviews in a market with non-Bayesian agents and a monopolist who makes the pricing decision; their analysis, based on a fluid approximation, is closely related to the one used in this paper.

Whereas the aforementioned papers focused on single-product settings, Pixton and Simchi-Levi (2020) studied the dynamics of social learning in the case of many products with substitutable demand. Using a fluid approximation, they proved that social learning makes the product substitution effects stronger; that is, consumers tend to buy and report reviews for incumbent products, leaving a new product underexplored for very long periods of time. They showed that asynchronous launch times may lead to doubly-exponential differences in market share. In our paper we assume that products are launched simultaneously and consumers incur search costs when browsing displayed items with rank, which introduces new challenges to the platform since the social learning transients can be influenced by the platform's ranking decision.

Mostly owing to the proliferation of online platforms, the analysis of information disclosure policies designed to incentivize consumers to learn about products has recently attracted quite a lot of attention. See, for instance, Frazier et al. (2014), Kremer et al. (2014), Papanastasiou et al. (2018), Bimpikis et al. (2019), Che and Hörner (2018). Whereas these papers focused on a platform's incentive mechanisms to facilitate learning, Zhao (2021) studied how a monopolistic platform can influence consumers' learning via product ranking. In her model, consumers form a consideration set consisting of some number of top-ranked products and then make a selection from the set following a choice model—hence, consumers' learning is influenced by product ranking. Zhao (2021) proposed an upper confidence bound (UCB) ranking algorithm that balances exploration-exploitation trade-offs and characterized its performance bounds. Golrezaei et al. (2021) studied the problem of learning product rankings when a platform faces a mixture of real and fake users. They proposed efficient learning algorithms and characterized their worst-case performance bounds. Our work, too, concerns

the platform's problem of learning product rankings, but differs from Zhao (2021) and Golrezaei et al. (2021) at least in two dimensions: our model assumes rank-dependent search cost; and our analysis focuses on a fluid formulation in which the performance of our proposed ranking policies can be expressed in semi-closed forms, rather than as bounds.

The product display problem analyzed in Section 4 belongs to the family of dynamic assortment optimization problems with multiple products under a general choice model. Starting with Talluri and Van Ryzin (2004), this type of problems has been investigated when the distribution of consumers' preferences is a priori known (Davis et al., 2014) and when preferences have to be learned by the designer along the selling horizon (Rusmevichientong et al., 2010, Sauré and Zeevi, 2013, Agrawal et al., 2019). Extant empirical works provided evidence of the ranking effects due to search cost in online market environments; see, e.g., Kempe and Mahdian (2008), Craswell et al. (2008), Lerman and Hogg (2014), and Ghose et al. (2013). In the context of search cost, L'Ecuyer et al. (2017) considered the static optimal ranking policy of a revenue-maximizing search engine. They investigated the platform's dilemma of balancing between minimizing expected consumers' regret and maximizing long-term profits. We use a similar approach with the additional complexity due to learning effects on the consumers' side. Abeliuk et al. (2016) studied the assortment optimization problem when consumers are influenced by the aggregate past purchases. They showed that a static ranking policy can be optimal despite the time-varying demand due to social influence. In a related paper Berbeglia et al. (2021) showed that a platform always benefits from market segmentation when consumers are shown a ranking of products relevant to their own segment, rather than showing a ranking of all products. Whereas these papers examined the platform's assortment/ranking problem when the platform is well informed about consumers' valuations of underlying products, our paper considers an uninformed platform that must learn them on the fly.

The platform's learning and ranking problem in the current paper is related to the literature of dynamic pricing when sellers do not know the demand function (see, e.g., Besbes and Zeevi, 2009, Broder and Rusmevichientong, 2012, den Boer and Zwart, 2014, Keskin and Zeevi, 2014). These studies highlighted the trade-off between exploration (learning) and exploitation (earning); in particular, most extant studies show poor performance of *myopic/greedy* policies that do not put explicit efforts in exploration and, as a consequence, suffer from a phenomenon called incomplete learning. Our paper shares an important common theme with these papers in that we rigorously characterize the performance loss incurred under a greedy ranking policy and modify it to guard against poor performance.

Notation. Throughout the paper, for any pair of functions $f(\cdot), g(\cdot)$, the notation $f(x) = O(g(x))$ indicates that there exists a positive constant M such that $f(x) \leq Mg(x)$ for $x \to \infty$, whereas $f(x) = \Omega(g(x))$ means that there exist two positive constants M such that $f(x) \geq Mg(x)$ for $x \to \infty$.

Moreover, $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$. We use "$\sim$" to denote asymptotic equivalence; formally, given functions $f(x)$ and $g(x)$, we define a binary relation $f(x) \sim g(x)$ as $x \to \infty$ if and only if $f(x)/g(x) \to 1$ as $x \to \infty$. Additionally, the terms decreasing and increasing are to be taken in the strict sense representing strictly increasing and strictly decreasing, respectively.

The organization of the paper. In Section 2 we introduce a demand model that incorporates review information and search cost, and formulate a platform's ranking problem. In Section 3 we provide a preliminary analysis of the consumer learning dynamics using a fluid approximation. In Section 4 we present our main theoretical results: we reformulate the platform's ranking problem using the aforementioned fluid approximation (Section 4.1); we characterize the optimal ranking policy in the full-information setting as a benchmark (Section 4.2); we propose the greedy policy and characterize the performance gap with the fully informed platform (Section 4.3); and we introduce the semi-greedy policy for performance improvement and examine its effectiveness (Section 4.4). In Section 5, we show via simulation studies the performance of the two ranking policies in a wide spectrum of scenarios. In Appendix A we examine several extensions to the basic settings. Proofs are collected in Appendix B.

2. Modeling Framework

2.1. Discrete Consumers Setting

Model overview. We consider a marketplace where a set of K substitutable goods or services—henceforth, called the *products*—are offered to a market of consumers who decide whether to buy one of them, or to choose a no-purchase option. For each $k \in \{1, \dots, K\}$, $q_k \in [0, 1]$ represents the intrinsic quality of product k and p_k is the fixed price of the product. The index $k = 0$ indicates the no-purchase option, which has known intrinsic quality $q_0 = 0$ and a price $p_0 = 0$.

Consumers are indexed by $n = 1, 2, \dots$, and arrive at random times $t_1, t_2, \dots$ according to a Poisson process with rate $\Lambda > 0$; they make a once-and-for-all decision, and never re-enter the market. Without loss of generality, we assume the normalization $\Lambda = 1$ throughout the paper. Initially, consumers do not know the quality of the products and, in order to make their purchase decision, use their available information to compute a vector of quality estimates $\hat{\boldsymbol{q}}_n := (\hat{q}_{1,n}, \dots, \hat{q}_{K,n})$, where $\hat{q}_{k,n}$ denotes the estimate of the quality of product k evaluated by consumer n.

In case consumer n decides to buy product k, she may have a positive or a negative experience with it; namely, consumer n' experience is $\nu_{k,n}$, where the $\nu_{k,n}$'s are i.i.d. binary variables that take values L (like) or D (dislike), with

$$\mathsf{P}(\nu_{k,n} = L) = 1 - \mathsf{P}(\nu_{k,n} = D) = q_k. \tag{2.1}$$

So, the quality q_k represents the probability that a buyer of product k gets a positive experience. A similar model was considered by Papanastasiou et al. (2018). In what follows, we formalize the functional relationships among demand, reviews, and product rankings.

Product rankings and search costs. The search cost associated with a given product depends on the position in which the product is displayed to consumers. To formalize the search cost as a function of product ranking, we let $\mathcal{Z}^K$ denote the set of all permutations of $\{1,\dots,K\}$ and let $\Delta(\mathcal{Z}^K)$ represent the space of all probability distributions over $\mathcal{Z}^K$. Elements of $\mathcal{Z}^K$ will be referred to as *position assignments* or, more simply, as *rankings*. Given a product ranking $\boldsymbol{z}=(z_1,\dots,z_K)\in\mathcal{Z}^K$, $z_k=j$ indicates that product k occupies the j-th highest position in the ranking. For instance, when $z_k=1$ ($z_k=K$), product k occupies the highest (lowest) position. In particular, if a product occupies the j-th highest position in the ranking, customers incur a search cost $g(j)$, where $g:\mathbb{N}\to\mathbb{R}$ is a strictly increasing function; without loss of generality, we normalize $g(1)=0$.

Purchase decision. Consider a class of i.i.d. random variables $(\alpha_{k,n})$, with $k=0,1,\dots,K$ and $n\in\mathbb{N}$. The random variable $\alpha_{k,n}$ represents the idiosyncratic preference of consumer n for product k. The distribution of $\alpha_{k,n}$ is common knowledge, but its realization is private information of consumer n. Given a vector of quality estimates $\hat{\boldsymbol{q}}_n$ and a position assignment $\boldsymbol{z}\in\mathcal{Z}^K$, consumer n assigns a utility $\alpha_{k,n}+\hat{q}_{k,n}-p_k-g(z_k)$ to the purchase of product k, and a utility of $\alpha_{0,n}+q_0-p_0=\alpha_{0,n}$ to the outside option. Then, she buys the product c_n that maximizes her estimated utility, i.e., $c_n=\arg\max_{k=0,1,\dots,K}\left\{\alpha_{k,n}+\hat{q}_{k,n}-p_k-g(z_k)\right\}$, where we let $g(z_0):=0$. Under the above assumptions, consumer n purchases product k with probability $d_k(\hat{\boldsymbol{q}}_n,\boldsymbol{z})$, hereafter referred to as the *demand function*; formally,

$$d_k(\hat{\boldsymbol{q}}_n,\boldsymbol{z}):=\mathsf{P}(c_n=k\mid\hat{\boldsymbol{q}}_n,\boldsymbol{z}). \tag{2.2}$$

We make the following assumption on the demand function.

Assumption 2.1. *For fixed K, there exists a positive constant $\delta<1$ such that for any $k\in\{1,\dots,K\}$, $d_k(\hat{\boldsymbol{q}},\boldsymbol{z})\geq\delta$ for any vector of quality estimates $\hat{\boldsymbol{q}}\in[0,1]^K$ and ranking $\boldsymbol{z}\in\mathcal{Z}^K$.*

The preceding assumption implies that the demand for each product is strictly positive (although possibly small) for any consumer, which ensures that a new review for each product will eventually enter the learning process. Assumption 2.1 is satisfied for well-known multinomial choice models with unbounded support for preference $\{\alpha_{k,n}\}$, including the multinomial logit (MNL) and nested logit models. For example, the MNL model postulates that consumers' preferences $\{\alpha_{k,n}:k=0,1,\dots,K\}$ have a standard Gumbel distribution, i.e., $\mathsf{P}(\alpha_{k,i}\leq x)=\exp\left(-\exp(-x)\right)$, where the demand function in (2.2) is written as

$$d_k(\hat{\boldsymbol{q}},\boldsymbol{z}):=\frac{\exp(\hat{q}_k-p_k-g(z_k))}{1+\sum_{j=1}^K\exp(\hat{q}_j-p_j-g(z_j))}. \tag{2.3}$$

In Section 3 we consider general demand functions that satisfy Assumption 2.1 for our analysis of learning transient. From Section 4 and on, we focus on the MNL demand function for our analysis of the platform's ranking problem to elucidate the key features of the problem in analytically tractable forms.

Review mechanism and information structure. Consumers who buy one product truthfully report online their experience. So consumer n, after purchasing product k, reports the online review $\nu_{k,n}$. The quantities $L_{k,n} \coloneqq \sum_{s=1}^{n-1} \mathbb{1}\{c_s = k \text{ and } \nu_{k,s} = L\}$ and $D_{k,n} \coloneqq \sum_{s=1}^{n-1} \mathbb{1}\{c_s = k \text{ and } \nu_{k,s} = D\}$, respectively, represent the numbers of positive and negative reviews for product k observed by consumer n, and $B_{k,n} \coloneqq L_{k,n} + D_{k,n}$ is the total number of purchases (and, consequently, of reviews) for product k observed by consumer n. The symbol $\boldsymbol{I}_n \coloneqq \{(L_{k,n}, D_{k,n}) : k = 1, \ldots, K\}$ denotes the whole information available to consumer n.

Quality estimation procedure. We assume that consumers use Bayesian updating for the unknown qualities and their common prior for product k is $\text{Beta}(L_{k,0}, D_{k,0})$.[1] This implies that the posterior belief over q_k is $\text{Beta}(L_{k,n} + L_{k,0}, D_{k,n} + D_{k,0})$. Hence, consumer n's Bayesian estimate of the quality of product k is given by

$$\hat{q}_{k,n} \coloneqq \frac{L_{k,n} + L_{k,0}}{L_{k,t} + L_{k,0} + D_{k,n} + D_{k,0}} = \frac{L_{k,n} + L_{k,0}}{B_{k,n} + B_{k,0}}. \tag{2.4}$$

As it is apparent in (2.4), $B_{k,0} \coloneqq L_{k,0} + D_{k,0}$ represents the weight that consumers assign to the prior belief $\hat{q}_{k,0}$.[2]

2.2. The Platform's Ranking Problem

The platform does not know $\boldsymbol{q}$ and receives a share $0 < \rho \leq 1$ of every payment that takes place on its website, i.e., the platform realizes a revenue ρp_k whenever product k is sold. Let $\boldsymbol{\sigma}_n = (\sigma_{1,n}, \ldots, \sigma_{K,n})$ be the position assignment observed by consumer n, where $\sigma_{k,n} = j$ indicates that product k is displayed in position j to consumer n. Consider the selling horizon of length $T > 0$ and let N_T be the index of the last consumer; although we consider a finite horizon T here, our main analyses in Sections 3 and 4 focus on performance metrics as functions of T (when T is sufficiently large). The platform commits to a non-anticipating[3] *ranking policy* $\boldsymbol{\Pi}_T \coloneqq \{\Pi_n : n = 1, \ldots, N_T\}$ which, given an information state $\boldsymbol{I}_n$ available to consumer n, returns a randomized ranking $\Pi_n \in \Delta(\mathcal{Z}^K)$, i.e., a probability distribution over the set of possible rankings $\mathcal{Z}^K$. More formally, for any consumer n, the probability distribution Π_n defines a vector $(\pi_{\boldsymbol{z}_1,n}, \pi_{\boldsymbol{z}_2,n}, \ldots, \pi_{\boldsymbol{z}_{K!},n})$, where $\pi_{\boldsymbol{z},n} \coloneqq \mathsf{P}(\boldsymbol{\sigma}_n = \boldsymbol{z})$ satisfies the obvious normalization constraint $\sum_{\boldsymbol{z} \in \mathcal{Z}^K} \pi_{\boldsymbol{z},n} = 1$.

Given the quality estimate $\hat{\boldsymbol{q}}_n$ of consumer n, the expected demand for product k under a randomized ranking Π_n is given by

$$\tilde{d}_k(\hat{\boldsymbol{q}}_n, \Pi_n) \coloneqq \sum\nolimits_{\boldsymbol{z} \in \mathcal{Z}^K} \pi_{\boldsymbol{z},n} d_k(\hat{\boldsymbol{q}}_n, \boldsymbol{z}). \tag{2.5}$$

The platform's objective is to choose a non-anticipating ranking policy that maximizes its expected cumulative revenue over a selling horizon. Formally, the platform's optimal control problem can be stated as follows:

$$\underset{\{\boldsymbol{\Pi}_T\}}{\text{maximize}} \quad \mathsf{E}_{\boldsymbol{q}} \left[\sum_{n=1}^{N_T} \sum_{k=1}^{K} \rho p_k \tilde{d}_k(\hat{\boldsymbol{q}}_n, \Pi_n) \right]. \tag{2.6}$$

It is quite difficult—if not impossible—to find an exact solution to the stochastic dynamic programming problem (2.6). If the true quality vector $\boldsymbol{q}$ were known, one could characterize the optimal solution to (2.6) via the Bellman equation, although obtaining it in closed form would be difficult. A more fundamental challenge arises from the fact the true quality vector $\boldsymbol{q}$ is unknown. In particular, the expected value in (2.6) is taken with respect to the unknown true quality vector $\boldsymbol{q}$, and hence the platform does not know how the current ranking would change beliefs and optimal ranking decisions for subsequent consumers.[4] In the following sections, we provide alternative methods to characterize consumers' learning transient in tractable form, which will be leveraged to derive structural insights into the platform's ranking problem.

Remark 2.1. In solving (2.6), the platform should learn the unknown quality while ultimately maximizing the accumulated revenue over time. Hence, the platform should judiciously balance the trade-off between learning and earning. In this regard, our ranking problem resembles combinatorial multi-armed bandit (CMAB) problems (Chen et al., 2013); specifically, for every customer arriving in the market, the platform has to choose an arm, i.e, a ranking $\boldsymbol{z} \in \mathcal{Z}^K$, and receives a reward given by the profit generated by the customer. However, there are fundamental differences between our ranking problem and CMAB problems: in CMAB problems, the expected reward for each arm is static but unknown, whereas in our ranking problem, the expected reward for each arm (ranking) is known but time-varying, depending on the previous ranking decisions. Specifically, the platform does not know how the demand function evolves over time (i.e., the ODE in (3.1)) because of the unknown quality vector $\boldsymbol{q}$. Thus, the methodologies used in CMAB problems cannot be directly applied to our setting.

3. Preliminaries: Fluid Approximation and Learning Transient

In this section we introduce a fluid model where the learning transients can be described as solutions of deterministic ordinary differential equation (ODE) systems, which is a significantly more tractable framework in many settings; see Crapis et al. (2017), Shin et al. (2021) for examples of recent applications of fluid models to revenue management problems in the presence of online reviews. In what follows, we will omit the detailed derivation of the fluid formulation, which can be found in Appendix C, and we only illustrate the intuition behind it.

3.1. Associated Fluid Approximation of the Learning Dynamics

For the sake of building insights, assume that consumers arrive continuously over time at a constant rate $\Lambda = 1$ so that over a small interval $[t, t + \mathrm{d}t)$ a mass $\mathrm{d}t$ of consumers enters the market. In a small interval $[t, t + \mathrm{d}t)$, the ranking $\boldsymbol{z} \in \mathcal{Z}^K$ is fixed and the state variables vary at a rate which is given by the expected variation of their discrete counterpart over $[t, t + \mathrm{d}t)$. That is, among the total

mass of $\mathrm{d}t$ consumers arriving in the market during the infinitesimal interval $[t, t+\mathrm{d}t)$, roughly a fraction $d_k(\hat{\boldsymbol{q}}_n, \boldsymbol{z})$ will purchase product k. The dynamics for the state variables $L_{k,n}$ and $D_{k,n}$ can be described in the same spirit as above, and the fluid approximation is obtained by taking limit $\mathrm{d}t \to 0$ in the system dynamics, along with a suitable scaling of the arrival rate Λ.

In what follows, we use the argument t in parenthesis to denote variables in continuous time; for example, $\hat{q}_k(t)$ represents consumers' quality estimate for product k at time t, which can be considered as counterparts of $\hat{q}_{k,n}$ for consumer n in the discrete consumers setting. Let $\boldsymbol{\sigma}(t) = (\sigma_1(t), \ldots, \sigma_K(t))$ be the position assignment observed by consumer t. The platform commits to a non-anticipating *ranking policy* $\boldsymbol{\Pi}(T) \coloneqq \{\Pi(t) : t \in [0,T]\}$ which, given an information state $\boldsymbol{I}(t)$ at any time $t \in [0,T]$, returns a randomized ranking $\Pi(t) \in \Delta(\mathcal{Z}^K)$, i.e., a probability distribution over the set of possible rankings $\mathcal{Z}^K$.

Along with the initial conditions $\hat{q}_k(0) = \hat{q}_{k,0}$ and $B_k(0) = 0$, the learning trajectories in the fluid approximation are governed by the following ODE system: for $k = 1, \ldots, K$,

$$\begin{cases} \dot{\hat{q}}_k(t) = \dfrac{\dot{B}_k(t)}{B_k(t) + B_{k,0}} \left[q_k - \hat{q}_k(t) \right], \\ \dot{B}_k(t) = \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi(t)), \end{cases} \tag{3.1}$$

and $\dot{L}_k(t) = q_k \dot{B}_k(t)$ and $\dot{D}_k(t) = (1 - q_k)\dot{B}_k(t)$. From the ODE, one can show that $\hat{q}_k(t)$ satisfies

$$\hat{q}_k(t) = \frac{L_k(t) + L_{k,0}}{B_k(t) + B_{k,0}}, \tag{3.2}$$

which resembles the quality estimate (2.4) in the discrete consumers setting. Notice that in (3.1) the time derivative of the number of purchases for product k is given by the expected value of the *ranking-dependent* demand function $\tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi(t))$ defined in (2.5). In contrast, the time derivative of the quality estimate $\hat{q}_k(t)$ is not dependent on the ranking policy $\Pi(t)$ since the latter has no direct impact on the consumers' belief updating procedure. Consequently, the ranking policy does not influence the direction of the change in $\hat{q}_k(t)$, but affects the speed of the learning process.

3.2. Structural Properties of the Learning Transient

Monotonicity and convergence. Using the fact that $L_k(t) = q_k B_k(t)$ in the fluid formulation, the quality estimate $\hat{q}_k(t)$ in (3.2) can be rewritten as

$$\hat{q}_k(t) = q_k - (q_k - \hat{q}_{k,0}) \frac{B_{k,0}}{B_{k,0} + B_k(t)}. \tag{3.3}$$

Therefore, the learning transient exhibits some monotonicity properties as is formalized in the following proposition.

PROPOSITION 3.1. *Consider the ODE system in (3.1). For any randomized ranking policy $\Pi(t) \in \Delta(\mathcal{Z}^K)$ for $t \geq 0$, $\hat{\boldsymbol{q}}(t) \to \boldsymbol{q}$ as $t \to \infty$. Furthermore,*

- *if* $\hat{q}_{k,0} = q_k$, *then* $\hat{q}_k(t) = q_k$ *for all* $t \geq 0$;
- *if* $\hat{q}_{k,0} < q_k$, *then* $\hat{q}_k(t)$ *is monotonically increasing for all* $t \geq 0$;
- *if* $\hat{q}_{k,0} > q_k$, *then* $\hat{q}_k(t)$ *is monotonically decreasing for all* $t \geq 0$.

To paraphrase the preceding proposition, consumers' quality estimates converge to the true quality vector $\boldsymbol{q}$ irrespective of the ranking policy adopted by the platform. Furthermore, consumers' perceived quality increases (decreases) if the true quality is higher (lower) than their initial estimate. These properties are useful in analyzing the speed of learning in the presence of product choice, which we discuss next.

Time-to-learn analysis. We now investigate how fast consumers' quality estimates converge to their limits and discuss how the speed of learning depends on various model primitives. To this end, we temporarily assume that there is no search cost (i.e., $g(\cdot) = 0$). For fixed $k \in \{1, \dots, K\}$, let us focus on the phase of the learning process for $t \leq \tau_k^K(\varepsilon)$, where, given a small positive constant ε,

$$\tau_k^K(\varepsilon) := \inf\{t > 0 : |\hat{q}_k(t) - q_k| \leq \varepsilon\} \tag{3.4}$$

is the ε-time-to-learn for product k when the market contains K products, which we use as a measure of the learning speed. Below, we provide insights on how $\tau_k^K(\varepsilon)$ depends on the market parameters relative to both product k and its competing products, which will be indicated with the index $j \neq k$. It is evident that the ε-time-to-learn depends on various model parameters: however, to limit the notation burden, we will not indicate this dependence explicitly in the definition of $\tau_k^K(\varepsilon)$.

In single-product settings, the ε-time-to-learn can be obtained in closed form; see, e.g., a similar discussion in Crapis et al. (2017). In multiproduct settings, $\tau_k^K(\varepsilon)$ does not admit a closed form owing to the substitution effect between products. However, we can make a number of qualitative statements about the ε-time-to-learn with respect to model primitives. We summarize these in the following theorem, where we use $S^K := \{(p_k, q_k, \hat{q}_{k,0}) : k = 1, \dots, K\}$ to describe a market with K products.

THEOREM 3.1. *Assume* $g(\cdot) := 0$. *Fix* $\varepsilon > 0$ *and assume that* $|q_k - \hat{q}_{k,0}| > \varepsilon$. *Let* $\tau_k^K(\varepsilon)$ *and* $\tau_k^{K+1}(\varepsilon)$ *be respectively the* ε*-time-to-learn for product* k *in the markets* S^K *and* S^{K+1} *such that* $S^{K+1} = S^K \cup \{(p_{K+1}, q_{K+1}, \hat{q}_{K+1,0})\}$. *Then,* $\tau_k^K(\varepsilon) < \tau_k^{K+1}(\varepsilon)$. *Moreover, for fixed* K, $\tau_k^K(\varepsilon)$ *is*

(i) *decreasing (increasing) in* $\hat{q}_{k,0}$ *and increasing (decreasing) in* q_k *if* $\hat{q}_{k,0} < q_k$ *(*$\hat{q}_{k,0} > q_k$*);*

(ii) *increasing in* q_j *and* $\hat{q}_{j,0}$ *for* $j \neq k$;

(iii) *increasing in* p_k *and* $B_{k,0}$, *and decreasing in* p_j *and* $B_{j,0}$ *for* $j \neq k$.

Based on the above result, we can provide several insights on the factors that influence the learning transient the most.

(a) *Number of product options*: The time-to-learn $\tau_k^K(\varepsilon)$ increases with K, the number of products,[5] because the demand for each product is diluted when a new product is added in the market. The effect of the new product on the time-to-learn is further analyzed in Proposition B.1 in Section B.1, where we provide a tight upper bound for $\tau_k^K(\varepsilon)$; for instance, when prices are all set equal to p, there exists a certain market condition in which

$$\tau_k^K(\varepsilon) \sim \frac{B_{k,0}}{\varepsilon}(\mathrm{e}^{p-1}+K) \quad \text{as} \quad \varepsilon \to 0. \tag{3.5}$$

In particular, $\tau_k^K(\varepsilon)$ increases linearly with K for a sufficiently small $\varepsilon > 0$, which is a consequence of the increased substitution effect when a new product is added in the market.

(b) *Prior belief*: The time-to-learn is increasing in the distance of the prior from the truth; for example, the more consumers initially underestimate (or overestimate) the true quality of a product, the longer it takes for consumers to learn the truth. Furthermore, the time-to-learn is increasing with $B_{k,0}$, the weight of prior estimate; that is, the higher the weight assigned to the prior belief, the larger the number of reviews required to forget prior estimates, thus slowing down learning.

(c) *Relative attractiveness vs. competing alternatives*: The time-to-learn for a product depends on how the product is attractive relative to other products. Specifically, Theorem 3.1 implies that $\tau_k^K(\varepsilon)$ increases with both $\hat{q}_{j,0} - p_j$ and $q_j - p_j$, i.e., the initial and the eventual attractiveness of its competing products $j \neq k$, respectively. In other words, more attractive products (either because they started from a higher prior belief, or because they have higher intrinsic quality, or because they are cheaper) will be selected more frequently by consumers, hindering information accumulation for their competitors.

4. The Platform's Ranking Problem: A Fluid Formulation

In the absence of search cost, Theorem 3.1 suggests that consumers' learning transients are correlated across products due to the substitution effects. In this section, we assume that the search cost is positive and strictly increasing with the position in the ranking. The interplay between the substitution and ranking effects makes the learning transients more complicated. To facilitate transparent analysis of the platform's ranking problem, we will focus on the MNL model throughout this section, where the demand function is given as (2.3).

4.1. Fluid Formulation under Multinomial Logit Demand

The platform's ranking problem in the fluid formulation. Recall from Theorem 3.1 that irrespective of the ranking policy adopted by the platform, consumers' quality estimates converge to the true quality $\boldsymbol{q}$ via social learning. As a result, the platform has no control over the asymptotic

learning outcome. Conversely, search cost can have a potentially significant impact on the learning speed since, by picking the product ordering, the platform can affect product choice and the speed of information acquisition. In particular, information acquisition for products placed in the highest positions is much faster with search cost, compared to the case without search cost. The opposite happens for products displayed in the lowest positions in the ranking: the platform may need a high number of iterations—possibly, exponential in the number of products—to discover the most profitable products in the market.

For the sake of building intuition, suppose that search cost increases linearly with the displayed position, i.e., $g(k) = \gamma(k-1)$, where γ is a positive constant. Consider two deterministic position assignments $\boldsymbol{z}$ and $\boldsymbol{z}'$, and suppose that $\boldsymbol{z}$ places product k exactly one position lower than $\boldsymbol{z}'$ does, that is, $z_k = z'_k + 1$. It is easy to see that, for large enough values of K and all other things being equal, $d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) \simeq \mathrm{e}^{-\gamma} d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}')$. Namely, the demand function of product k roughly decreases of a factor $\mathrm{e}^{-\gamma}$ when the ranking of product k is decreased by exactly one unit. In other words, the demand, and hence the learning speed, of the product at position k is roughly $\mathrm{e}^{-\gamma k}$ times smaller than that of the top-ranked product. Concretely, if $\gamma = 0.2$, then customers effectively restrict their option set to the first 15-20 products, whereas if $\gamma = 0.8$, then this is true only for the top 4-6 products.

The platform does not know $\boldsymbol{q}$ and receives a share $0 < \rho \leq 1$ of every payment that takes place on its website, i.e., the platform realizes a revenue ρp_k whenever product k is sold. The platform's objective is to choose a non-anticipating ranking policy that maximizes its expected cumulative revenue over a selling horizon of length $T > 0$. Formally, given a quality configuration $Q = (\boldsymbol{q}, \boldsymbol{q}_0) \in \mathcal{Q}^K$, the platform's optimal control problem can be stated as follows:

$$\begin{aligned} R_T^*(Q) := \underset{\{\boldsymbol{\Pi}(T)\}}{\text{maximize}} \quad & \mathsf{E}_{\boldsymbol{q}} \left[\int_0^T \sum_{k=1}^K \rho p_k \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi(t)) \, \mathrm{d}t \right] \\ \text{subject to} \quad & \text{ODE in (3.1)}. \end{aligned} \tag{4.1}$$

Notice that the expected value in (4.1) is taken with respect to the unknown true quality vector $\boldsymbol{q}$. That is, even though the fluid model approximation removes the discreteness and stochasticity of consumer demand and the heterogeneity of consumers' preferences and of the ex-post quality noise, enabling a deterministic description of the learning transients, the platform still faces a stochastic control problem with respect to the unknown true quality vector $\boldsymbol{q}$, which affects the learning dynamics and the achievable revenue objective.

4.2. Full Information Benchmark and the Notion of Regret

The oracle platform. Define $\boldsymbol{z}_\infty$ as the position assignment that maximizes revenues if consumers make decisions based on $\boldsymbol{q}$:

$$\boldsymbol{z}_\infty \coloneqq \operatorname*{arg\,max}_{\boldsymbol{z}\in\mathcal{Z}^K} \left\{ \sum_{k=1}^{K} p_k d_k(\boldsymbol{q}, \boldsymbol{z}) \right\}.$$

For simplicity, it will be assumed that $\boldsymbol{z}_\infty$ is unique in the remainder of the paper.[6] The optimal revenue rate r_∞ is defined accordingly as

$$r_\infty = \sum_{k=1}^{K} p_k d_k(\boldsymbol{q}, \boldsymbol{z}_\infty). \tag{4.2}$$

The following proposition characterizes the optimal policy for the control problem in (4.1) for an *oracle* platform that knows the true quality vector $\boldsymbol{q}$. In this full-information benchmark, the oracle platform's ranking decision may convey information about the product's quality, but we assume that consumers do not adjust their quality estimate in response to that information.

PROPOSITION 4.1. *If the platform knows $\boldsymbol{q}$, then, there exists a unique solution $\Pi^*(t)$ to the platform optimal control problem (4.1). Moreover, there exists $T_0 < \infty$ such that for all $T \geq T_0$, $\Pi^*(t)$ satisfies $\pi_{\boldsymbol{z}_\infty}(t) = 1$ for all $t \in [0, T]$.*

In the preceding proposition, the condition $T \geq T_0$ ensures that the time horizon is sufficiently large such that under the optimal policy, the true ranking can be recovered at the end of the selling horizon[7]; that is, $\hat{q}_{k_1}(T) \leq \hat{q}_{k_2}(T)$ if $q_{k_1} \leq q_{k_2}$ for any $k_1 \neq k_2$. Note that the threshold T_0 depends on the search cost $g(\cdot)$; for example, in the case of linear search cost $g(k) = \gamma(k-1)$, the demand for the product at the kth position is roughly of order $\mathrm{e}^{-\gamma k}$, so T_0 must be of order $\mathrm{e}^{\gamma K}$ to ensure sufficient time to learn the qualities of all products. Proposition 4.1 establishes that, if the selling horizon is large enough, then it is optimal for the platform to adopt a static deterministic ranking policy that displays products according to the asymptotically optimal ranking $\boldsymbol{z}_\infty$ throughout $[0, T]$. The optimality of Π^* guarantees that the revenue achieved by an *oracle* platform that knows $\boldsymbol{q}$ (and, hence, $\boldsymbol{z}_\infty$), henceforth denoted by R_T^*, provides an upper bound for the revenue achieved by any other policy Π implemented without knowing the optimal ranking $\boldsymbol{z}_\infty$.

Proposition 4.1 highlights another important aspect of the interplay between this profit maximizing platform and the consumer learning process. Specifically, since the optimal policy is static and displays products according to the asymptotically optimal ranking $\boldsymbol{z}_\infty$, for an uninformed platform it is beneficial to design ranking policies that allow to discover $\boldsymbol{z}_\infty$ as quickly as possible. This suggests that the platform and the consumers have aligned interests, as they both have a strong incentive to discover $\boldsymbol{q}$ in the shortest amount of time. Specifically, if consumers knew $\boldsymbol{q}$, they would be able to choose the product that, given their personal preferences and the price, best fits their needs, whereas, if the platform knew $\boldsymbol{q}$, it would use this information to derive $\boldsymbol{z}_\infty$ and achieve the optimal revenue rate r_∞.

The regret of a ranking policy. The performance metric we will use throughout this study is the *long-run regret*, defined as

$$\mathfrak{R}_\Pi(Q) := \lim_{T\to\infty} \left\{ R_T^*(Q) - R_T^\Pi(Q) \right\}, \tag{4.3}$$

where, given a configuration $Q := (\boldsymbol{q}, \hat{\boldsymbol{q}}_0)$, $R_T^*(Q)$ is the T-period optimal revenue achieved by the oracle platform characterized in Proposition 4.1 under Q, and $R_T^\Pi(Q)$ is the T-period revenue achieved by the ranking policy Π under Q. Note that the regret $\mathfrak{R}_\Pi(Q)$ can be infinite for some policies whose ranking does not converge to $\boldsymbol{z}_\infty$ in the long run. However, it will be shown later that the regret is finite for reasonable ranking policies: among such policies our analysis will focus on *asymptotically optimal* ranking policies such that $\pi_{\boldsymbol{z}_\infty}(t) \to 1$ as $t \to \infty$.

Instead of solving the stochastic dynamic programming problem (4.1), throughout the paper we adopt a worst-case scenario approach to the ranking optimization problem, that is, we take on the challenge of identifying the market circumstances that are most adversarial to the ranking policy adopted by the platform. The worst-case analysis provides an important piece of information for a more complete evaluation of the platform's ranking policy. In particular, we aim at identifying the configuration $Q \in \mathcal{Q}^K$ that maximizes the regret $\mathfrak{R}_\Pi(Q)$. Formally, given the ranking policy $\Pi \in \Delta(\mathcal{Z}^K)$, the worst-case regret is denoted by

$$\mathfrak{R}_\Pi^K := \max_{Q\in\mathcal{Q}^K} \{\mathfrak{R}_\Pi(Q)\}. \tag{4.4}$$

The maximizer of (4.4) is denoted by Q_Π^K, henceforth referred to as the worst-case configuration under the ranking policy Π over K products.

4.3. Regret Analysis for the Greedy Policy

The greedy ranking policy. In online marketplaces, the ranking decision must be made on a real-time basis. The platform, uninformed of the true quality vector, may not be able to efficiently solve the stochastic dynamic programming problem (4.1) but rather employs a computationally tractable solution. In this section, we focus on the *greedy* policy where the platform makes the ranking decision to maximize the instantaneous revenue rate as if the current estimates of the qualities were accurate. In general contexts of dynamic programming that involve learning, this type of policy is often dismissed by practitioners since it does not acquire sufficient information about unknown features of the model, incurring a significant loss in revenue (see, e.g., den Boer and Zwart, 2014, Keskin and Zeevi, 2014). In our problem, such a policy does not suffer from incomplete learning because the true quality of each product will be eventually revealed to the market, even though the platform does not exert an explicit effort to explore the product quality (Theorem 3.1). As will be shown below, this policy passes a basic sanity check in that the regret (4.3) is finite in the long run.

Formally, at any $t \geq 0$, the *greedy ranking policy* G displays a position assignment drawn from the probability distribution $\Pi^{\mathsf{G}}(t) \in \Delta(\mathcal{Z}^K)$ such that

$$\Pi^{\mathsf{G}}(t) = \operatorname*{arg\,max}_{\Pi \in \Delta(\mathcal{Z}^K)} \mathsf{E}_{\Pi}\left[\sum_{k=1}^{K} p_k \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi) \middle| \hat{\boldsymbol{q}}(t)\right] = \operatorname*{arg\,max}_{\Pi \in \Delta(\mathcal{Z}^K)} \sum_{k=1}^{K} p_k \sum_{\boldsymbol{z} \in \mathcal{Z}^K} \pi_{\boldsymbol{z}}\, \tilde{d}_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}), \tag{4.5}$$

where E_{Π} denotes the expected value with respect to the probability distribution Π. In fact, solving (4.5) is equivalent to finding the solution of a combinatorial optimization problem over the space of possible deterministic position assignments $\mathcal{Z}^K$ (Lemma C.2); formally, for $t \geq 0$, we have

$$\mathsf{E}_{\Pi^{\mathsf{G}}(t)}\left[\sum_{k=1}^{K} p_k\, \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi^{\mathsf{G}}(t))\right] = \max_{\boldsymbol{z} \in \mathcal{Z}^K} \sum_{k=1}^{K} p_k\, d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}). \tag{4.6}$$

Specifically, under the greedy policy, $\Pi^{\mathsf{G}}(t)$ is a degenerate probability distribution that assigns maximal probability to the position assignment $\boldsymbol{z}^{\mathsf{G}}(t) = (z_1^{\mathsf{G}}(t), \ldots, z_K^{\mathsf{G}}(t))$, which is the solution of the combinatorial optimization problem on the right-hand side of (4.6), hereafter referred to as the *Multinomial Logit Positioning Problem* (MNLPP).

Even if the number of permutations of $\{1, \ldots, K\}$ grows super-exponentially with K, Abeliuk et al. (2015) showed that MNLPP can be solved in polynomial time, and that any optimal position assignment $\boldsymbol{z}^{\mathsf{G}}(t)$ for (4.6) and the corresponding optimal revenue $r^{\mathsf{G}}(t) := \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}^{\mathsf{G}}(t))$ are such that

$$z_{k_1}^{\mathsf{G}}(t) \leq z_{k_2}^{\mathsf{G}}(t) \quad \Longleftrightarrow \quad (p_{k_1} - r^{\mathsf{G}}(t))\mathrm{e}^{\hat{q}_{k_1}(t) - p_{k_1}} \geq (p_{k_2} - r^{\mathsf{G}}(t))\mathrm{e}^{\hat{q}_{k_2}(t) - p_{k_2}}, \tag{4.7}$$

for all $k_1, k_2 \in \{1, \ldots, K\}$. Notice that if there exist $k_1 \neq k_2$ such that $p_{k_1} = p_{k_2}$, then (4.7) implies that only estimated qualities matter for determining the relative position of products k_1 and k_2, i.e., $z_{k_1}^{\mathsf{G}}(t) \leq z_{k_2}^{\mathsf{G}}(t)$ if and only if $\hat{q}_{k_1}(t) \geq \hat{q}_{k_2}(t)$. In particular, if prices are all equal, then $\boldsymbol{z}^{\mathsf{G}}(t)$ simply displays products in decreasing order of their current estimated quality, i.e.,

$$p_k = p \text{ for all } k \in \{1, \ldots, K\} \quad \Longrightarrow \quad z_k^{\mathsf{G}}(t) = \sum_{j=1}^{K} \mathbb{1}\{\hat{q}_k(t) \leq \hat{q}_j(t)\} \text{ for all } k \in \{1, \ldots, K\}. \tag{4.8}$$

We assume that in case of ties the platform ranks products in alphabetic order, i.e., when there exist two products $k_1 \neq k_2$ such that $p_{k_1} = p_{k_2}$ and $\hat{q}_{k_1}(t) = \hat{q}_{k_2}(t)$ then $z_{k_1}^{\mathsf{G}} < z_{k_2}^{\mathsf{G}}$ if and only if $k_1 < k_2$.

Worst-case regret for the greedy policy. The analysis of this section will be conducted under the following assumption.

ASSUMPTION 4.1. (a) *$B_{k,0} = B_0$ for all $k = 1, \ldots, K$.*

(b) *$p_k = p$ for all $k = 1, \ldots, K$.*

(c) *$\hat{q}_{1,0} \leq q_k$ for each $k \neq 1$.*

Assumption 4.1(a) is needed for clarity of exposition. Assumption 4.1(b) simplifies the analysis, but it has no bearing on our qualitative insights. It is appropriate for products such as smartphone apps and movies, whose price is typically fixed at a level common to the industry. Moreover, the assumption is approximately true for quality-differentiated products for which customers are more sensitive to quality than price. Our extended analysis for the case of different prices is provided in Section B.2. Finally, Assumption 4.1(c) ensures that the prior estimate of the best product (i.e., product 1) is sufficiently low, so that it is initially ranked low under the greedy policy; we focus on this setting that is more adversarial to the greedy policy than the one with high $\hat{q}_{1,0}$. As we will see, this assumption is not restrictive in our worst-case analysis since $\hat{q}_{1,0}$ is equal to zero in the worst case for sufficiently large K (Proposition 4.2).

For worst-case analysis, it will be useful to focus first on the set of configurations $\mathcal{Q}^K(\eta) \subset \mathcal{Q}^K$, which, given some constant $\eta \in (0,1)$, is defined as

$$\mathcal{Q}^K(\eta) := \left\{ (\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \middle| q_1 \le 1 \quad \text{and} \quad q_1 - \eta = q_2 \ge q_3 \ge \cdots \ge q_K \ge 0, \quad \hat{\boldsymbol{q}}_0 \in [0,1]^K \right\}, \tag{4.9}$$

that is, $\mathcal{Q}^K(\eta)$ contains configurations where the highest quality is greater than the second-highest quality by η. Note that $\mathcal{Q}^K = \bigcup_{\eta \in (0,1)} \mathcal{Q}^K(\eta)$.

For fixed $\eta \in (0,1)$, let $\mathfrak{R}_{\mathsf{G}}^K(\eta) := \sup\{\mathfrak{R}_{\mathsf{G}}(Q) : Q \in \mathcal{Q}^K(\eta)\}$ be the worst-case regret under the greedy policy. The worst-case configuration, if it exists, is difficult to characterize precisely, but can be approximated when the number of products is sufficiently large, as is formalized in the next proposition.

To make this asymptotic analysis with respect to K precise, we need to define *nested markets*. Recall the definition of market $S^K := \{(p_k, q_k, \hat{q}_{k,0}) : k = 1, \dots, K\}$.

Definition 4.1 (Nested market). For any $K \ge 1$, we say that S^K is nested in S^{K+1} if $S^{K+1} = S^K \cup \{(p_{K+1}, q_{K+1}, \hat{q}_{K+1,0})\}$.

The nested market structure ensures that the consecutive markets differ only by one product, which allows us to capture marginal effects when one more product is added to the market. For any quality configuration with K products $Q^K := (\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \in [0,1]^K \times [0,1]^K$, for the sake of convenience the qualities will be reordered nonincreasingly: $q_1 \ge q_2 \ge \cdots \ge q_K$. We let $\mathcal{Q}^K$ be the set of all possible configurations Q^K.

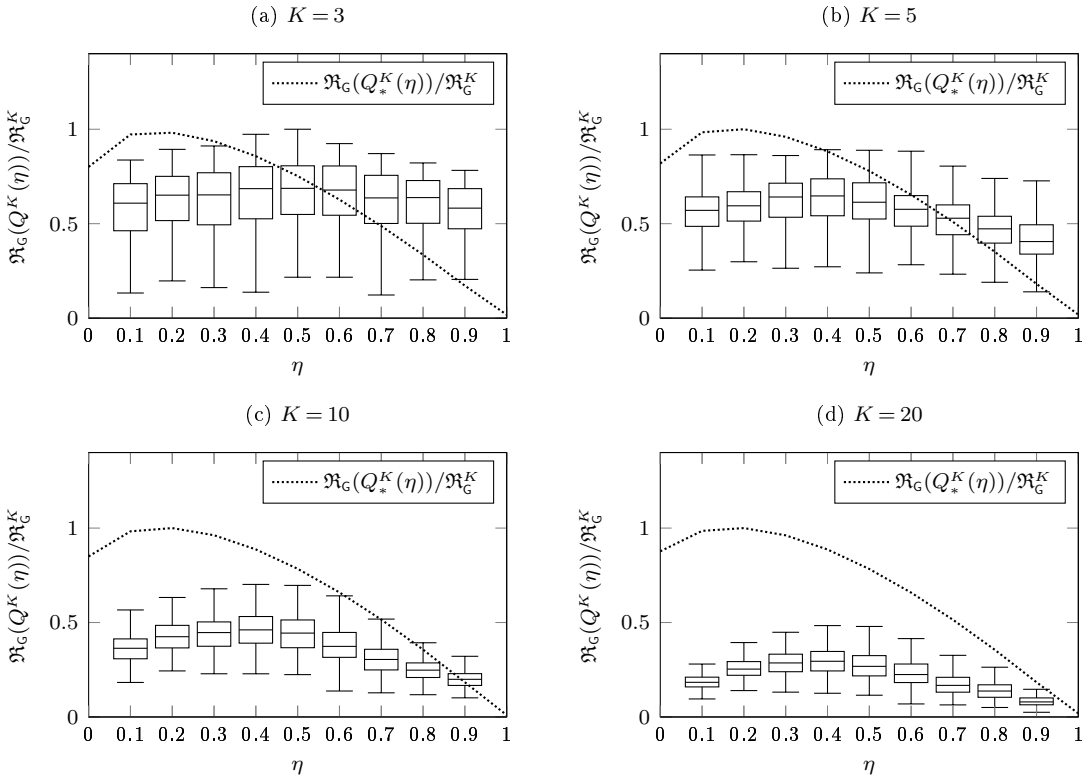
Proposition 4.2. *Fix $\eta \in (0,1)$ and suppose that Assumption 4.1 holds. Then, as $K \to \infty$,*

$$\frac{\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta))}{\mathfrak{R}_{\mathsf{G}}^K(\eta)} \to 1, \tag{4.10}$$

where $Q_*^K(\eta) := (\boldsymbol{q}^*, \hat{\boldsymbol{q}}_0^*)$ is such that

$$\begin{aligned} \boldsymbol{q}^* &= (1, 1-\eta, \dots, 1-\eta) \\ \hat{\boldsymbol{q}}_0^* &= (0, 1-\eta, \dots, 1-\eta). \end{aligned} \tag{4.11}$$

Figure 1 The regret of the greedy policy under randomly generated configurations.

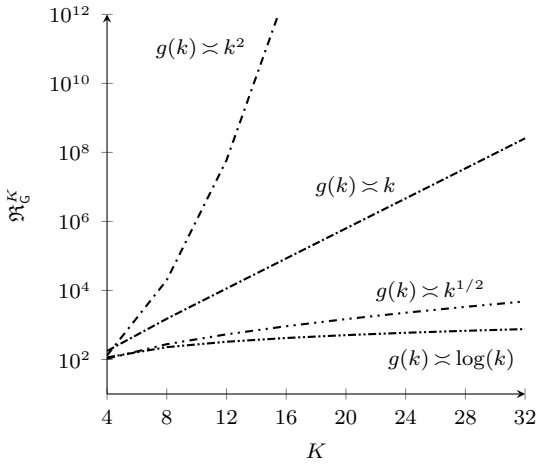


Note. The normalized regret, $\mathfrak{R}_{\mathsf{G}}(Q^K(\eta))/\mathfrak{R}_{\mathsf{G}}^K$, is plotted as a function of $\eta = q_1 - q_2$. The box plot for each η shows the minimum, maximum, median, and the first and third quartiles of the normalized regret over 300 randomly generated configurations.

The preceding proposition implies that the worst-case regret for the greedy policy is achieved, approximately, when (i) the best-quality product has the lowest-possible prior expectation (i.e., $q_1^* = 1$ and $\hat{q}_{1,0}^* = 0$), (ii) customers have perfect prior on the non-best products (i.e., $q_k^* = \hat{q}_{k,0}^*$ for $k \neq 1$), and (iii) the qualities of the non-best products are as high as possible (i.e., $q_k = 1 - \eta$ for $k \neq 1$). Therefore, when the number of products is sufficiently large, the worst-case regret for the greedy policy reduces to that of configuration $Q_*^K(\eta)$.

This result is illustrated in Fig. 1. Specifically, for each $\eta \in \{0.1, \ldots, 0.9\}$ we randomly generate 300 configurations $Q^K(\eta) \in \mathcal{Q}^K(\eta)$ as follows: for the quality, set $q_1 = 1$ and $q_2 = 1 - \eta$, and draw $q_k \sim \text{Beta}(3, 1)$ for $k \neq 1, 2$; and for the prior belief, set $q_{1,0} = 0$ and draw $q_{k,0} \sim \text{Uniform}(0, 1)$ for $k \neq 1$. In all cases, we assume a linear search cost $g(k) = 0.5k$. For a small number of products ($K = 3$), observe that $Q_*^K(\eta)$ may not be a worst-case configuration; that is, $\mathfrak{R}_{\mathsf{G}}(Q^K(\eta)) \geq \mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta))$ for some configuration $Q^K(\eta)$. For a large number of products ($K = 20$), $Q_*^K(\eta)$ is a worst-case configuration for which $\mathfrak{R}_{\mathsf{G}}(Q^K(\eta)) \leq \mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta))$ for any random configuration $Q^K(\eta)$.

Figure 2 The growth of the worst-case regret of the greedy policy.



Note. The worst-case regret $\mathfrak{R}_{\mathsf{G}}^{K}$ is plotted as a function of K in log-linear scale. In all cases, the regret is estimated over the same selling horizon $[0,T]$ for a large T.

In configuration $Q_*^K(\eta)$, the regret can be characterized in a tractable form, which is given in the theorem that we now present. Before stating the result, for $x_1, x_2 < 0$, we define the function ψ as

$$\psi(x_1,x_2) \coloneqq \int_{x_1}^{x_2} \frac{\mathrm{e}^{-y}}{y^2}\,\mathrm{d}y. \tag{4.12}$$

THEOREM 4.1. *Fix $\eta \in (0,1)$. Under Assumption 4.1, consider the (asymptotic) worst-case configuration $Q_*^K(\eta)$ characterized in Proposition 4.2. Then, as $K \to \infty$,*

$$\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta)) \sim \mathfrak{M}_{\mathsf{G}}^K(\eta) \coloneqq \frac{\left(\mathrm{e}^{1-p} - \mathrm{e}^{1-p-\eta}\right)\left(\mathrm{e}^{g(K)}-1\right)}{1+\mathrm{e}^{1-p}+\mathrm{e}^{1-p-\eta}\sum_{j=2}^{K}\mathrm{e}^{-g(j)}}\psi(-1,-\eta), \tag{4.13}$$

where $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ is non-negative, continuous, and quasi-concave in $\eta \in [0,1]$ with $\mathfrak{M}_{\mathsf{G}}^K(0) = \mathfrak{M}_{\mathsf{G}}^K(1) = 0$.

The dotted lines in Fig. 1 illustrate the (normalized) regret $\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta)) \sim \mathfrak{M}_{\mathsf{G}}^K(\eta)$ as a function of η. The quasi-concavity of $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ implies that the regret $\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta))$ is maximized at a unique $\eta^* = \arg\max_{\eta\in[0,1]}\{\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta))\}$ for sufficiently large K. Combined with Proposition 4.2, the preceding theorem provides an approximate characterization of the worst-case regret (4.4) under the greedy policy, which is formalized in the following corollary.

COROLLARY 4.1. *Under Assumption 4.1, the worst-case regret under the greedy policy satisfies*

$$\mathfrak{R}_{\mathsf{G}}^K = \Theta\left(\frac{\mathrm{e}^{g(K)}}{\sum_{k=1}^{K}\mathrm{e}^{-g(k)}}\right) \quad \text{as} \quad K \to \infty. \tag{4.14}$$

As alluded to earlier, the greedy policy puts little emphasis on learning and more on exploiting profit. Therefore, although the greedy policy achieves a finite regret for given K, the platform may incur significant loss of revenue as the number of competing products grows large. In particular, Corollary 4.1 indicates how fast the revenue loss in the worst case increases with K due to such under-exploratory behavior. Concretely, consider the following examples:

- (Linear cost function) If $g(k) = \gamma(k-1)$ for some $\gamma > 0$, then the worst-case regret increases at an *exponential* rate with K; that is, $\mathfrak{R}_{\mathsf{G}}^{K} = \Theta(\mathrm{e}^{K})$.
- (Logarithmic cost function) If $g(k) = \gamma \log(k)$ for some $\gamma > 0$, then the worst-case regret increases at most at a *linear* rate with K; that is, $\mathfrak{R}_{\mathsf{G}}^{K} = \Theta(K/\log(K))$.

These observations are generalized in the following corollary.

COROLLARY 4.2. *Suppose that Assumption 4.1 holds. If $g(k) = \Omega(k^{\alpha})$ for some $\alpha > 0$, then the regret grows exponentially with K; formally,*

$$\liminf_{K\to\infty} \frac{\mathfrak{R}_{\mathsf{G}}^{K+1}}{\mathfrak{R}_{\mathsf{G}}^{K}} > 1. \tag{4.15}$$

If $g(k) = O(k^{\alpha})$ for any $\alpha > 0$, then the regret does not grow exponentially with K; formally,

$$\limsup_{K\to\infty} \frac{\mathfrak{R}_{\mathsf{G}}^{K+1}}{\mathfrak{R}_{\mathsf{G}}^{K}} = 1. \tag{4.16}$$

Corollaries 4.1 and 4.2 suggest that the platform should consider the implications of the search cost on the design of the ranking system. Several comments are in order.

First, our findings identify the parametric regimes of market environments that are favorable (and not favorable) to the greedy policy. Specifically, if the platform wishes the regret to be subexponential in K, the greedy policy can be a desirable solution only when the search cost is subpolynomial (Corollary 4.2); in such circumstances, even if the platform places a "good" product at a low rank, consumers sufficiently explore such a product and *help* the platform eventually raise its rank. In contrast, if consumers incur significant search cost that is polynomial with product position, then the platform essentially *shades* the low-ranked products from consumers (i.e., preempts the consumers' opportunity to learn), thereby suffering from an exponential revenue loss as the number of product increases. In such circumstances, an important managerial implication is that the greedy policy alone cannot be a desirable approach and the platform should consider, for instance, a market segmentation strategy (Berbeglia et al., 2021): instead of showing a ranking of all products, the platform may segment consumer population and show a ranking of products that are relevant to a specific segment.

Furthermore, firms display a different number of products to customers on different platforms. For example, mobile phones have smaller screens than do PCs, which increases the cognitive cost associated with information gathering (Ghose et al., 2013). Thus, it is typical to display a smaller number of products on mobile versions of the platform than on PC versions (e.g., Amazon and Netflix). Our analysis provides a rough guideline for choosing how many products to display. Specifically, if the platform employs the greedy policy and aims to achieve the regret due to the ranking effect less than a constant $C > 0$, then the platform should display $O(g^{-1}(\log(C)))$ products to customers. Concretely, in the case of the linear cost function, the desired number of products is $O(\log(C))$, whereas in the case of the logarithmic cost function, the desired number of products is $O(C)$.

4.4. Regret Analysis for the Semi-greedy Ranking Policy

The semi-greedy ranking policy. Our regret analysis of the greedy ranking policy indirectly emphasizes the value of *exploration*, as it suggests that the platform has a strong incentive to discover the high-quality products especially when the growth of search cost is polynomial with the product position. To improve the (worst-case) performance of the greedy policy, we now consider the *semi-greedy* ranking policy, denoted by SG, which is structured around the *implied* belief process $\{\tilde{q}_k(t) : t > 0\}$ defined as

$$\tilde{q}_k(t) := \hat{q}_k(t) + \frac{u}{B_k(t) + B_{k,0}}, \tag{4.17}$$

where $u > 0$ is the parameter of the policy that controls the level of exploration;[8] note that the implied belief $\tilde{q}_k(t)$ may exceed one. It is easily seen that

$$\dot{\tilde{q}}_k(t) = \frac{\dot{B}_k(t)}{B_k(t) + B_{k,0}}(q_k - \tilde{q}_k(t)), \tag{4.18}$$

which means that the implied belief process $\tilde{q}_k(t)$ is increasing (decreasing) at t if $\tilde{q}_k(t) < q_k$ (if $\tilde{q}_k(t) > q_k$). For each t, the semi-greedy policy ranks the product based on the implied belief $\tilde{\boldsymbol{q}}(t) = (\tilde{q}_1(t), \dots, \tilde{q}_K(t))$; that is,

$$\boldsymbol{z}^{\mathsf{SG}} \in \arg\max_{\boldsymbol{z} \in \mathcal{Z}^K} \left\{ \sum_{k=1}^{K} p_k d_k(\tilde{\boldsymbol{q}}(t), \boldsymbol{z}) \right\}. \tag{4.19}$$

The above problem can be framed as another instance of MNLPP, whose optimal solution can be characterized as in (4.7), with $\hat{q}_k(t)$ being replaced with $\tilde{q}_k(t)$. Because of the additional term in (4.17), the semi-greedy policy puts more emphasis on exploration than the greedy; for instance, although the estimated quality of product k is lower than that of k' (i.e., $\hat{q}_k(t) < \hat{q}_{k'}(t)$), product k can be ranked higher if it is under-explored compared to product k' (i.e., $B_k(t) < B_{k'}(t)$).

Worst-case regret for the semi-greedy policy. For fixed $\eta \in [0,1]$, let $\mathfrak{R}^K_{\mathsf{SG}}(\eta) := \sup\{\mathfrak{R}_{\mathsf{SG}}(Q) : Q \in \mathcal{Q}^K(\eta)\}$ denote the worst-case regret under the semi-greedy policy. The following proposition suggests that the worst-case scenarios for the semi-greedy and greedy policies coincide, although the corresponding regrets may differ significantly.

THEOREM 4.2. *Fix $\eta \in (0,1)$. Under Assumption 4.1, consider the semi-greedy policy parametrized by a positive constant $u < \bar{u} := B_0(1-\eta)$. Then, $\mathfrak{R}^K_{\mathsf{SG}}(\eta) \sim \mathfrak{R}_{\mathsf{SG}}(Q^K_*(\eta))$ as $K \to \infty$, where $Q^K_*(\eta)$ is characterized in (4.11). Furthermore, as $K \to \infty$,*

$$\mathfrak{R}_{\mathsf{SG}}(Q^K_*(\eta)) \sim \mathfrak{M}^K_{\mathsf{SG}}(\eta) := \frac{(\mathrm{e}^{1-p} - \mathrm{e}^{1-p-\eta})\left(\mathrm{e}^{g(K)} - 1\right)}{1 + \mathrm{e}^{1-p} + \mathrm{e}^{1-p-\eta}\sum_{j=2}^{K} \mathrm{e}^{-g(j)}} \psi\left(-1, -\frac{\eta B_0}{B_0 - u}\right), \tag{4.20}$$

where $\mathfrak{M}^K_{\mathsf{SG}}(\eta)$ is non-negative, continuous, and quasi-concave in $\eta \in [0,1]$ with $\mathfrak{M}^K_{\mathsf{SG}}(0) = \mathfrak{M}^K_{\mathsf{SG}}(1) = 0$ and ψ is defined in (4.12). Additionally, $\mathfrak{M}^K_{\mathsf{SG}}(\eta) < \mathfrak{M}^K_{\mathsf{G}}(\eta)$, where $\mathfrak{M}^K_{\mathsf{G}}(\eta)$ is defined in Theorem 4.1.

In the preceding theorem, the condition $u \leq \bar{u}$ ensures that the estimated quality of the best product is initially low such that $\tilde{q}_1(0) < q_k$ for $k \neq 1$. As in Assumption 4.1(b), this condition is imposed to ensure that the worst-case regret is characterized in a tractable form, but we remark that the semi-greedy policy admits the parameter $u > \bar{u}$ in general settings. Theorem 4.2 implies that, as in the greedy policy, the worst-case regret under the semi-greedy policy grows with K in the order equivalent to $\mathrm{e}^{g(K)}/\sum_{k=1}^{K}\mathrm{e}^{-g(k)}$. However, it is important to note that the worst-case regret under the semi-greedy policy increases at a slower rate than under the greedy; namely, $\mathfrak{M}_{\mathsf{SG}}^{K}(\eta) < \mathfrak{M}_{\mathsf{G}}^{K}(\eta)$. Several comments on this relation are in order.

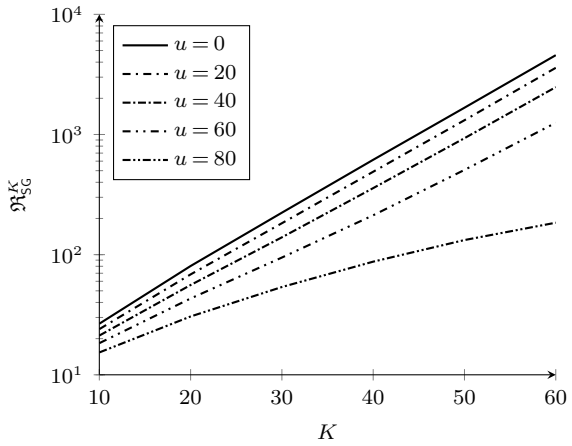
The comparison of Theorems 4.1 and 4.2 suggests that the semi-greedy policy can significantly reduce the worst-case regret by changing the exploration-exploitation balance; specifically, whereas the greedy policy gives priority to the product with high estimated quality, the semi-greedy policy puts more emphasis on learning the quality of products that are less explored. Concretely, recall from Theorems 4.1 and 4.2 that the worst-case regrets $\mathfrak{R}_{\mathsf{G}}^{K}(\eta)$ and $\mathfrak{R}_{\mathsf{SG}}^{K}(\eta)$ for the greedy and semi-greedy policies are approximated by $\mathfrak{M}_{\mathsf{G}}^{K}(\eta)$ and $\mathfrak{M}_{\mathsf{SG}}^{K}(\eta)$, respectively, which satisfy, for any $K \geq 2$ and $u \in (0, \bar{u})$,

$$\frac{\mathfrak{M}_{\mathsf{SG}}^{K}(\eta)}{\mathfrak{M}_{\mathsf{G}}^{K}(\eta)} = \frac{\psi\left(-1, -\frac{\eta B_0}{B_0 - u}\right)}{\psi\left(-1, -\eta\right)} < 1, \tag{4.21}$$

where the function ψ is the exponential integration function defined in (4.12). This ratio represents the efficiency of the semi-greedy policy relative to the greedy; in particular, as the ratio is close to zero, the semi-greedy policy is considered more efficient relative to the greedy. The ratio is equal to one for $u = 0$ (in which case the greedy and semi-greedy policies are identical) and is decreasing with $u \in [0, \bar{u}]$. In other words, the performance of the semi-greedy policy improves as the level of exploration, u, increases in the worst-case configuration. However, note that the preceding arguments are made based on the worst-case analysis, and it may not be always beneficial to increase the level of exploration u in general configurations; see Section 5 for the effect of u on the performance of the semi-greedy policy for general configurations.

Fig. 3 illustrates these observations when the (linear) search cost function is given as $g(k) = 0.1k$. The figure depicts the worst-case regret of the semi-greedy policy as a function of K for different values of the parameter $u \in \{0, 20, 40, 60, 80\}$, where $\bar{u} > 80$ in all cases. As is anticipated by Theorems 4.1 and 4.2, one can observe that under the semi-greedy policy, the worst-case regret $\mathfrak{R}_{\mathsf{SG}}^{K}$ is of order e^{K}, or, equivalently, $\log(\mathfrak{R}_{\mathsf{SG}}^{K})$ is approximately linear in K. Surprisingly, as a simple modification from the greedy, the semi-greedy policy can reduce the regret by orders of magnitude, highlighting the benefit of balancing between exploration and exploitation.

Figure 3 The growth of the worst-case regret of the semi-greedy policy.



Note. The worst-case regret $\mathfrak{R}^K_{\mathsf{SG}}$ is plotted as a function of K in log-linear scale. In all cases, $p_k = 1$ and $B_{k,0} = 100$ for all $k = 1, \dots, K$, and the search cost function is $g(k) = 0.1k$. Note that the semi-greedy policy with $u = 0$ is identical with the greedy.

5. Numerical Analysis

Whereas our theoretical analysis in Section 4 focuses on the worst-case regret of the proposed policies, this section numerically investigates the regret in a wide spectrum of scenarios for the platform's ranking problem.

Benchmark policy. In addition to the greedy and semi-greedy policies discussed in Section 4, we consider the *explore-then-exploit* ranking policy, denoted by EtE, which consists of two stages: an initial systematic exploration stage, where products are sequentially displayed in the top-position until a sufficient number of reviews are accumulated for each product, and then a full exploit stage, where the platform myopically chooses the ranking to maximize the immediate revenue rate. The EtE ranking policy is parametrized by $\bar{B}$, the minimum number of purchases for each product at the end of the exploration stage, which is summarized as follows.

- (Exploration) The exploration stage consists of K phases indexed by $i = 1, \dots, K$. For phase i, products are displayed according to $\boldsymbol{z}^{\mathsf{EtE}}(t) := (z_1^{\mathsf{EtE}}(t), \dots, z_K^{\mathsf{EtE}}(t))$, where

$$z_k^{\mathsf{EtE}}(t) = \begin{cases} k - i + K + 1 & \text{if } \ k < i, \\ k - i + 1 & \text{if } \ k \geq i, \end{cases} \tag{5.1}$$

  until $B_i(t) \leq \bar{B}$. Whenever $B_i(t) = \bar{B}$, the policy moves to the next phase $i+1$. Note that phase $i+1$ can be skipped if $B_{i+1}(t) > \bar{B}$ at the end of phase i.
- (Exploitation) After the exploration stage, products are displayed according to $\boldsymbol{z}^{\mathsf{G}}(t)$ by the greedy policy.

Notice that after product i is displayed in the top position during phase i, it occupies the last position in the $(i+1)$th phase, the second last in the $(i+2)$th phase, etc. The idea behind the EtE

Table 1 Regret under linear search cost.

$\mathfrak{R}_\Pi(Q^K)$	$K=10$			$K=50$		
	G	SG	EtE	G ($\cdot 10^4$)	SG ($\cdot 10^4$)	EtE ($\cdot 10^4$)
Min.	0.00	0.00	26.37	0.02	0.12	0.25
1%	0.03	7.49	54.90	0.31	0.15	0.33
25%	36.95	31.10	107.45	50.33	0.18	0.49
50%	98.07	55.09	150.45	390.60	0.22	0.62
75%	216.06	100.93	215.62	1969.81	0.33	0.86
99%	859.49	315.96	492.73	49576.21	2.56	3.80
Max.	1731.92	912.81	1343.02	394392.12	69.70	25.68
Average	158.42	77.06	173.95	3286.82	0.37	0.82
Std. dev.	182.22	67.87	93.20	12098.21	0.88	0.81

Notes. In all cases, we set $g(k) = 0.5(k-1)$. The numbers are summary statistics of the regret calculated in 10^5 random scenarios.

policy is to guarantee a sufficiently long exploration phase for each product by displaying it at the top position, so that in the exploitation stage, the estimated quality of each product is not too far from the true quality.

Experimental settings. In this numerical study, we fix $p_k = 1$ and $B_{k,0} = 100$ for all k and consider randomly generated quality configurations. Specifically, the quality q_k is generated from a uniform distribution on $[0,1]$. Then, the prior belief $\hat{q}_{k,0}$ is generated from $\text{Beta}(a_k, b_k)$, where a_k and b_k are chosen such that $\mathsf{E}[\hat{q}_{k,0}] = q_k$ and $\text{Var}[\hat{q}_{k,0}] = 0.4^2$. Note that Assumption 4.1(c) is relaxed in our numerical study. We consider two types of search cost: the linear search cost $g(x) = 0.5(x-1)$ and the logarithmic search cost $g(x) = \log(x)$. For each $K \in \{10, 50\}$, we consider 10^5 randomly generated quality configurations. For each configuration Q^K, we calculate the regret $\mathfrak{R}_\Pi(Q^K)$ for each policy $\Pi \in \{\mathsf{G}, \mathsf{SG}, \mathsf{EtE}\}$. The performance of the EtE and SG policies depend on the tuning parameters $\bar{B}$ and u, respectively. For fair comparison of these policies, for each K, we calculate the regret for different tuning parameters $\bar{B} \in \{20, 40, \ldots, 200\}$ and $u \in \{20, 40, \ldots, 100\}$ and choose the ones that give the smallest median regret among the 10^5 random configurations: in the case of linear cost, we choose $(\bar{B}, u) = (20, 60)$ for $K = 10$ and $(\bar{B}, u) = (180, 80)$ for $K = 50$; and in the case of logarithmic cost, we choose $(\bar{B}, u) = (20, 60)$ for $K = 10$ and $(\bar{B}, u) = (20, 20)$ for $K = 50$. These are by no means optimal choices in general circumstances, but we have found that the key qualitative conclusions do not depend on these choices.

Results and discussion. In the case of linear search cost, the estimated values of the regret are summarized in Table 1. As anticipated by Theorems 4.1 and 4.2, the regrets under the greedy and semi-greedy policies grow exponentially with K in scenarios that are near the worst case; in particular, for the 99th percentile, the regret for $K = 50$ is more than 10^3 times greater than that for $K = 10$ for both greedy and semi-greedy policies. Although both policies exhibit exponential growth, the semi-greedy policy significantly reduces the growth rate of the regret; in particular, as K increases from 10 to 50, the regret under the greedy policy increases by a factor of 10^4 in

Table 2 Regret under logarithmic search cost.

$\mathfrak{R}_\Pi(Q^K)$	$K=10$			$K=50$		
	G	SG	EtE	G	SG	EtE
Min.	0.00	0.00	11.80	7.17	14.04	177.22
1%	0.00	1.65	25.16	31.12	38.15	213.47
25%	8.10	10.47	47.21	90.65	91.50	278.23
50%	21.62	18.40	58.64	136.10	131.56	312.98
75%	50.34	35.69	76.94	200.42	188.38	357.77
99%	171.12	105.09	148.49	444.45	400.88	517.67
Max.	255.67	166.83	223.19	711.72	637.03	677.04
Average	36.37	26.92	63.63	156.55	149.26	323.96
Std. dev.	40.12	23.82	27.27	90.62	79.82	64.88

Notes. In all cases, we set $g(k)=\log(x)$. The numbers are summary statistics of the regret calculated in 10^5 random scenarios.

the median case, whereas the regret under the semi-greedy policy increases only by a factor of 40. For overall scenarios, the semi-greedy policy judiciously balances the trade-off between exploration and exploitation, and thus exhibits robust performance compared to the greedy. Note that the EtE policy also exhibits robust performance across the wide spectrum of scenarios. However, since EtE blindly puts all products into exploration, even some product whose quality is obviously low, its performance is poor relative to the semi-greedy policy (except for some extreme cases).

In the case of logarithmic search cost, the estimated values of the regret are reported in Table 2. Compared to the circumstances with linear search cost, low ranked products are only moderately penalized, so that the performance of the greedy policy is not severely bad. Concretely, recall from Corollary 4.2 that the regret grows only linearly with K under the greedy policy in the worst case. In contrast to the case of linear search cost, where the regret increases by orders of magnitude as K increases from 10 to 50, one can observe from Table 2 that the regret is comparable between the cases with $K=10$ and $K=50$. Since the greedy policy is favorable in this market environment with logarithmic search cost, the semi-greedy policy does not make a significant improvement from the greedy. The EtE policy, however, performs poorly in most scenarios; because of the (relatively) low search cost, there is less need of forced exploration, making the EtE policy overly conservative.

Endnotes

1. Recall that, the probability density function g_{Beta} of a beta with shape parameters a,b is given by

$$g_{\text{Beta}}(x) := \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^a (1-x)^b, \qquad x \in [0,1], \tag{5.2}$$

where $\Gamma(z)$ is the *gamma* function.

2. Although our model focuses on the Bayesian updating based on the beta-Bernoulli pair, this assumption has no bearing on our results. Our mathematical analysis can be easily extend to

commonly used prior-posterior conjugate pairs such as normal-normal, beta-binomial, and gamma-Poisson pairs.

3. We say that $\mathbf{\Pi}_T = \{\Pi_n : n = 1, \dots, N_T\}$ is non-anticipating if Π_n is only allowed to depend on past information $\boldsymbol{I}_n$.

4. The impact of partial information regarding the underlying demand function has been studied in detail in the revenue management literature in the setting where the demand function is unknown but constant over time; see, e.g., den Boer and Zwart (2014), Keskin and Zeevi (2014). In contrast, in our problem setting, the demand function itself evolves over time in conjunction with the perceived qualities of the offered products. Although the demand function is known at each time point, the platform cannot anticipate how it evolves over time because of the lack of information about the products' true quality.

5. The quantities $\underline{w}_k$ and $\bar{w}_k$ depend on K through the $K-1$ (strictly positive) summands in $\sum_{j\neq k} \exp(q_j - p_j)$ and $\sum_{j\neq k} \exp(\hat{q}_{j,0} - p_j)$ respectively.

6. To guarantee that $\boldsymbol{z}_\infty$ is unique, it suffices to assume that when there are ties between products, it is optimal for the platform to rank products, for instance, in alphabetic order, i.e., $z_{k_1} < z_{k_2}$ iff $k_1 < k_2$.

7. For instance, recalling (3.4), if ε is small enough, so that $\varepsilon \leq |q_{k_1} - q_{k_2}|$ for all $k_1, k_2 = 1, \dots, K$, then this assumption holds when $T \geq \max_k \tau_k^K(\varepsilon)$.

8. Our analysis easily extends to the case where the parameter u depends on the product index k, but we suppress the dependence to simplify analysis and exposition.

Acknowledgments

Stefano Vaccari gratefully acknowledges the hospitality of DRO, Columbia Business School of New York, where a good part of this research was carried out. Marco Scarsini is a member of GNAMPA-INdAM. This research project received partial support from the COST action GAMENET, the Italian MIUR PRIN 2017 Project ALGADIMAR "Algorithms, games, and digital markets," and the GNAMPA-INdAM 2020 grant "Random walks on random games." We thank Alberto Marcati for some relevant references.

References

Abeliuk A, Berbeglia G, Cebrian M, Van Hentenryck P (2015) The benefits of social influence in optimized cultural markets. *PloS ONE* 10(4):1–20.

Abeliuk A, Berbeglia G, Cebrian M, Van Hentenryck P (2016) Assortment optimization under a multinomial logit model with position bias and social influence. *4OR* 14(1):57–75.

Acemoglu D, Makhdoumi A, Malekian A, Ozdaglar A (2017) Fast and slow learning from reviews. Technical Report 24046, National Bureau of Economic Research, URL http://dx.doi.org/10.3386/w24046.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-Bandit: a dynamic learning approach to assortment selection. *Oper. Res.* 67(5):1453–1485.

Banerjee AV (1992) A simple model of herd behavior. *Quart. J. Econom.* 797–817.

Berbeglia F, Berbeglia G, Van Hentenryck P (2021) Market segmentation in online platforms. *Eur. J. Oper. Res.* 295(3):1025–1041.

Besbes O, Scarsini M (2018) On information distortions in online ratings. *Oper. Res.* 66(3):597–610.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Political Econom.* 100(5):992–1026.

Bimpikis K, Ehsani S, Mostagir M (2019) Designing dynamic contests. *Oper. Res.* 67(2):339–356.

Bressan A, Piccoli B (2007) *Introduction to the Mathematical Theory of Control*, volume 2 of *AIMS Series on Applied Mathematics* (American Institute of Mathematical Sciences (AIMS), Springfield, MO).

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.

Cesari L (1966) Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I. *Trans. Amer. Math. Soc.* 124:369–412.

Che YK, Hörner J (2018) Recommender systems as mechanisms for social learning. *Quart. J. Econom.* 133(2):871–925.

Chen N, Li A, Talluri K (2021) Reviews and self-selection bias with operational implications. *Management Sci.* forthcoming.

Chen W, Wang Y, Yuan Y (2013) Combinatorial multi-armed bandit: General framework and applications. *Internat. Conf. Machine Learn.*, 151–159 (PMLR).

Crapis D, Ifrach B, Maglaras C, Scarsini M (2017) Monopoly pricing in the presence of social learning. *Management Sci.* 63(11):3586–3608.

Craswell N, Zoeter O, Taylor M, Ramsey B (2008) An experimental comparison of click position-bias models. *Proc. 2008 Internat. Conf. Web Search and Data Mining*, 87–94 (ACM).

Davis JM, Gallego G, Topaloglu H (2014) Assortment optimization under variants of the nested logit model. *Oper. Res.* 62(2):250–273.

den Boer AV, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* 60(3):770–783.

Feldman P, Papanastasiou Y, Segev E (2019) Social learning and the design of new experience goods. *Management Sci.* 65(4):1502–1519.

Frazier P, Kempe D, Kleinberg J, Kleinberg R (2014) Incentivizing exploration. *Proc. 15th ACM Conf. Econom. Comput.*, 5–22 (New York, NY, USA: Association for Computing Machinery).

Ghose A, Goldfarb A, Han SP (2013) How is the mobile internet different? search costs and local activities. *Inform. Systems Res.* 24(3):613–631.

Golrezaei N, Manshadi V, Schneider J, Sekar S (2021) Learning product rankings robust to fake users. *Proc. 22nd ACM Conf. Econom. and Comput.*, 560–561.

He QC, Chen YJ (2017) Dynamic pricing of electronic products with consumer reviews. *Omega* 80:123–134.

Ifrach B, Maglaras C, Scarsini M, Zseleva A (2019) Bayesian social learning from consumer reviews. *Oper. Res.* 67(5):1209–1221.

Kakhbod A, Lanzani G, Xing H (2021) Heterogeneous learning in product markets. Technical report, SSRN, URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3961223.

Kempe D, Mahdian M (2008) A cascade model for externalities in sponsored search. Papadimitriou C, Zhang S, eds., *Proc. Internet and Network Econom.: WINE 2008*, 585–596 (Springer Berlin Heidelberg).

Keskin NB, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* 62(5):1142–1167.

Kremer I, Mansour Y, Perry M (2014) Implementing the "wisdom of the crowd". *J. Political Econom.* 122(5):988–1012.

Kurtz TG (1977/78) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Appl.* 6(3):223–240.

L'Ecuyer P, Maillé P, Stier-Moses NE, Tuffin B (2017) Revenue-maximizing rankings for online platforms with quality-sensitive consumers. *Oper. Res.* 65(2):408–423.

Lerman K, Hogg T (2014) Leveraging position bias to improve peer recommendation. *PloS ONE* 9(6):e98914.

Papanastasiou Y, Bimpikis K, Savva N (2018) Crowdsourcing exploration. *Management Sci,* 64(4):1727–1746.

Papanastasiou Y, Savva N (2017) Dynamic pricing in the presence of social learning and strategic consumers. *Management Sci.* 63(4):919–939.

Pixton C, Simchi-Levi D (2020) Network effects, customer reviews, and product proliferation in online durable goods markets. Technical report, SSRN 3593773, URL http://dx.doi.org/10.2139/ssrn.3593773.

Rusmevichientong P, Shen ZJM, Shmoys DB (2010) Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Oper. Res.* 58(6):1666–1680.

Sauré D, Zeevi A (2013) Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Oper. Management* 15(3):387–404.

Shin D, Vaccari S, Zeevi A (2021) Dynamic pricing with online reviews. Technical report, SSRN eLibrary, URL https://ssrn.com/abstract=3439836.

Shin D, Zeevi A (2021) Product quality and information sharing in the presence of reviews. Technical report, SSRN eLibrary, URL https://ssrn.com/abstract=3886954.

Stenzel A, Wolf C, Schmidt P (2020) Pricing for the stars. dynamic pricing in the presence of rating systems. Technical report, Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim.

Stigler GJ (1961) The economics of information. *J. Political Econom.* 69(3):213–225.

Talluri K, Van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. *Management Sci.* 50(1):15–33.

Yang N, Zhang R (2018) Dynamic pricing and inventory management in the presence of online reviews. Technical report, SSRN eLibrary, URL https://ssrn.com/abstract=2571705.

Zhao J (2021) *Learning to Optimize Decisions in Online Service Platforms*. Ph.D. thesis, Columbia University, URL https://doi.org/10.7916/d8-jg1m-3a03.

Appendix A: List of Symbols

The following table contains the symbols that have been used throughout the paper.

$B_{k,n}$	$L_{k,n} + D_{k,n}$
$B_k(t)$	amount of purchases of product k in the interval $[0,t]$ (in the fluid approximation)
c_n	product that maximizes consumer n's expected utility
$d_k(\hat{\boldsymbol{q}}_n, \boldsymbol{z})$	$\mathsf{P}(c_n = k \mid \hat{\boldsymbol{q}}_i, \boldsymbol{z})$, defined in (2.2)
$\tilde{d}_k(\hat{\boldsymbol{q}}_n, \Pi_n)$	$\sum_{\boldsymbol{z}\in\mathcal{Z}^K} \pi_{\boldsymbol{z},n} d_k(\hat{\boldsymbol{q}}_n, \boldsymbol{z})$, defined in (2.5)
$D_{k,n}$	$\sum_{s=1}^{n-1} \mathbb{1}\{c_s = k \text{ and } x_{k,s} = D\}$
$D_k(t)$	amount of unfavorable reviews for product k in the interval $[0,t]$ (in the fluid approximation)
g	search cost function
G	greedy ranking policy
$\boldsymbol{I}_n$	$\{(L_{k,n}, D_{k,n}) : k = 1, \dots, K\}$, whole information available to consumer n
$\boldsymbol{I}(t)$	information state at time t (in the fluid approximation)
k	product index
K	number of products
$L_{k,n}$	$\sum_{s=1}^{n-1} \mathbb{1}\{c_s = k \text{ and } x_{k,s} = L\}$
$L_k(t)$	amount of favorable reviews for product k in the interval $[0,t]$ (in the fluid approximation)
$\mathfrak{M}_{\mathsf{G}}^K(\eta)$	$\frac{(\mathrm{e}^{1-p} - \mathrm{e}^{1-p-\eta})\left(\mathrm{e}^{g(K)}-1\right)}{1+\mathrm{e}^{1-p}+\mathrm{e}^{1-p-\eta}\sum_{j=2}^{K}\mathrm{e}^{-g(j)}} \psi(-1,-\eta)$, defined in (4.13)
n	consumer index
N_T	index of the last customer in a selling horizon of length $T > 0$
p_k	price of product k
q_k	quality of product k
$\hat{q}_{k,n}$	estimated quality of product k evaluated by consumer n
$\hat{\boldsymbol{q}}_n$	$(\hat{q}_{1,n}, \dots, \hat{q}_{K,n})$
$\hat{q}_k(t)$	estimated quality of product k at time t (in the fluid approximation)
Q^K	$(\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \in [0,1]^K \times [0,1]^K$, quality configuration with K products
Q_Π^K	maximizer of (4.4)
$\mathcal{Q}^K$	set of all possible configurations Q^K
$\mathcal{Q}^K(\eta)$	defined in (4.9)
r_∞	$\sum_{k=1}^K p_k d_k(\boldsymbol{q}, \boldsymbol{z}_\infty)$, defined in (4.2)
R_T^*	revenue achieved by the oracle platform
$R_T^\Pi(Q)$	T-period revenue achieved by the ranking policy Π under Q
$\mathfrak{R}_\Pi(Q)$	$\lim_{T\to\infty}\{R_T^*(Q) - R_T^\Pi(Q)\}$, long-run regret, defined in (4.3)
$\mathfrak{R}_\Pi^K$	$\max_{Q\in\mathcal{Q}^K}\{\mathfrak{R}_\Pi(Q)\}$, defined in (4.4)
S^K	$\{(p_k, q_k, \hat{q}_{k,0}) : k = 1, \dots, K\}$, market with K products
SG	semi-greedy ranking policy
t	time (continuous)
T	horizon
u	level of exploration
$z_k = j$	product k occupies the j-th highest position in the ranking
$\boldsymbol{z}_\infty$	$\arg\max_{\boldsymbol{z}\in\mathcal{Z}^K}\left\{\sum_{k=1}^K p_k d_k(\boldsymbol{q}, \boldsymbol{z})\right\}$, defined in (4.2)
$\boldsymbol{z}$	$(z_1, \dots, z_K) \in \mathcal{Z}^K$
$\mathcal{Z}^K$	set of all permutations of $\{1, \dots, K\}$
$\boldsymbol{z}^{\mathsf{G}}(t)$	$(z_1^{\mathsf{G}}(t), \dots, z_K^{\mathsf{G}}(t))$, the solution of $\max_{\boldsymbol{z}\in\mathcal{Z}^K}\sum_{k=1}^K p_k\, d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z})$, in (4.6)
0	index of the outside option

$\alpha_{k,n}$	consumer n's preference for product k
δ	positive constant such that, $d_k(\hat{\boldsymbol{q}}, \boldsymbol{z}) \geq \delta$, defined in Assumption 2.1
$\Delta(\mathcal{Z}^K)$	space of all probability distributions over $\mathcal{Z}^K$
Λ	rate of the Poisson process
$\nu_{k,n}$	consumer n's experience after buying product k
$\pi_{\boldsymbol{z},n}$	$\mathsf{P}(\boldsymbol{\sigma}_n = \boldsymbol{z})$
Π_n	$(\pi_{\boldsymbol{z_1},n}, \pi_{\boldsymbol{z_2},n}, \dots, \pi_{\boldsymbol{z_{K!}},n})$
$\boldsymbol{\Pi}_T$	$\{\Pi_n : n = 1, \dots, N_T\}$, ranking policy
$\Pi(t)$	randomized ranking at time t (in the fluid approximation)
$\boldsymbol{\Pi}(T)$	$\{\Pi(t) : t \in [0,T]\}$, ranking policy (in the fluid approximation)
$\Pi^{\mathsf{G}}(t)$	$\arg\max_{\Pi \in \Delta(\mathcal{Z}^K)} \mathsf{E}_\Pi \left[\sum_{k=1}^K p_k \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi) \mid \hat{\boldsymbol{q}}(t)\right]$, defined in (4.5)
ρ	platform's share of every payment that takes place on its website
$\rho^{\mathsf{G}}(t)$	optimal profit
$\sigma_{k,n} = j$	product k is displayed in position j to consumer n
$\boldsymbol{\sigma}(t)$	$(\sigma_1(t), \dots, \sigma_K(t))$, position assignment observed at time t
$\boldsymbol{\sigma}_n$	$(\sigma_{1,n}, \dots, \sigma_{K,n})$, position assignment observed by consumer n
τ_n	random arrival time of consumer n
$\tau_k^K(\varepsilon)$	$\inf\{t > 0 : \lvert\hat{q}_k(t) - q_k\rvert \leq \varepsilon\}$, ε-time-to-learn for product k when the market contains K products, defined in (3.4)
$\psi(x_1, x_2)$	$\int_{x_1}^{x_2} \mathrm{e}^{-y}/y^2 \, \mathrm{d}y$, defined in (4.12)

Appendix B: Additional Theoretical Results

B.1. Worst-Case Analysis for the Time-to-Learn

In the absence of search cost, we characterize the configuration under which consumers experience the longest time to learn a product's quality. For fixed $\varepsilon \in (0,1)$, this quality configuration can be identified as the solution of

$$\max_{Q \in \mathcal{Q}^K} \{\tau_k^K(\varepsilon) \text{ s.t. } \lvert\hat{q}_{k,0} - q_k\rvert \leq \varepsilon\}, \tag{B.1}$$

where $\mathcal{Q}^K$ represents the set of admissible quality configurations defined as

$$\mathcal{Q}^K := \left\{ (\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \middle| \begin{array}{c} 1 \geq q_1 \geq q_2 \geq \cdots \geq q_K \geq 0 \\ \hat{\boldsymbol{q}}_0 \in [0,1]^K \end{array} \right\}. \tag{B.2}$$

Proposition B.1. *Fix $\varepsilon \in (0,1)$ and assume that $\lvert q_k - \hat{q}_{k,0}\rvert > \varepsilon$. Consider the fluid model approximation in (3.1). Then, ε-time-to-learn for product k is maximized under quality configuration $Q_k^K = (\boldsymbol{q}, \hat{\boldsymbol{q}}_0)$ such that $\boldsymbol{q} = (1, \dots, 1)$ and $\hat{q}_{k,0} = 0$ and $\hat{q}_{j,0} = 1$ for $j \neq k$. In this configuration, we have that*

$$\tau_k^K(\varepsilon) = B_{k,0} \left(\frac{1-\varepsilon}{\varepsilon} + \psi(-1, -\varepsilon) \mathrm{e}^{p_k - 1} \left(1 + \sum\nolimits_{j \neq k} \mathrm{e}^{1-p_j} \right) \right). \tag{B.3}$$

Proof of Proposition B.1. Using Theorem 3.1, we know that the time-to-learn is maximized when $\hat{q}_{j,0} = q_j = 1$ for $j \neq k$ and $\lvert q_k - \hat{q}_{k,0}\rvert$ is maximal. Hence, we only have to check the quality configurations $\hat{q}_{k,0} = 0$, $q_k = 1$ and $\hat{q}_{k,0} = 1$, $q_k = 0$ to find the maximum time-to-learn. Call these two

configurations A, and B, and let $\tau_k^{K,A}$ and $\tau_k^{K,B}$ be the ε-time-to-learn for A and B respectively; in this proof, we fix ε and suppress it in function arguments for a clear exposition.

Consider configuration A initially, and notice that, for all $j \in \{1,\dots,K\}$, $\hat{q}_{j,0} = q_j$ implies $\hat{q}_{j,0} = \hat{q}_j(t) = q_j$ for all $0 \le t \le \tau_k^{K,A}$. Then, defining $C_{-k} := 1 + \sum_{j\neq k} \exp(q_j - p_j) = 1 + \sum_{j\neq k} \exp(1 - p_j)$, we have

$$\begin{aligned} B_k(\tau_k^{K,A}) &= B_{k,0}\frac{1-\varepsilon}{\varepsilon} \\ &= \int_0^{\tau_k^{K,A}} \dot{B}_k(t)\,\mathrm{d}t = \int_0^{\tau_k^{K,A}} \frac{\exp(\hat{q}_k(t) - p_k)}{1 + \sum_{j\neq k}\exp(q_j - p_j) + \exp(\hat{q}_k(t) - p_k)}\,\mathrm{d}t \\ &= \tau_k^{K,A} - C_{-k}\int_0^{\tau_k^{K,A}} \frac{\mathrm{d}t}{1 + \sum_{j=1}^{K}\exp(\hat{q}_j(t) - p_j)}, \end{aligned}$$

where the first equality follows from (3.3) and the fact that $\hat{q}_k(\tau_k^{K,A}) = q_k - \varepsilon = 1 - \varepsilon$. Using Lemma C.1 with $g = 0$, this leads to

$$\tau_k^{K,A} = B_{k,0}\frac{1-\varepsilon}{\varepsilon} + C_{-k}B_{k,0}\exp(p_k - 1)\psi(-\varepsilon, -1),$$

where ψ is defined as in (4.12). Using a similar logic, we can obtain

$$\tau_k^{K,B} = B_{k,0}\frac{1-\varepsilon}{\varepsilon} + C_{-k}B_{k,0}\exp(p_k)[-\psi(\varepsilon, 1)].$$

Notice that the desired result follows if we prove $\tau_k^{K,B} - \tau_k^{K,B} > 0$, which is equivalent to show that

$$\exp(-1)\psi(-\varepsilon,-1) + \psi(\varepsilon,1) = \int_1^{\varepsilon} \frac{\exp(y-1) - \exp(-y)}{y^2}\,\mathrm{d}y > 0.$$

To prove the above inequality, observe that there exist a constant $c > 0$ such that

$$\begin{aligned} \int_1^{\varepsilon} \frac{\exp(y-1) - \exp(-y)}{y^2}\,\mathrm{d}y &\ge c\int_1^{\varepsilon} \exp(y-1) - \exp(-y)\,\mathrm{d}y \\ &= c[1 + \exp(-1) - \exp(\varepsilon - 1) - \exp(-\varepsilon)]. \end{aligned}$$

It is not difficult to show that the last term in the above series of inequalities is positive for every $\varepsilon \in (0,1)$, which proves $\tau_k^{K,B} - \tau_k^{K,B} > 0$ and concludes the proof. □

Proposition B.1 states that $\tau_k^K(\varepsilon)$ is maximized for a quality configuration under which product k is the least attractive product (in terms of estimated quality) throughout its learning transient. In this worst-case scenario, the value of the intrinsic quality of product k is the highest possible ($q_k = 1$), but consumers initially estimate its quality at the lowest possible level ($\hat{q}_{k,0} = 0$). In other words, the initial estimation bias $q_k - \hat{q}_{k,0}$ for product k is maximal. At the same time, in this worst-case scenario for product k, the estimated qualities of all the $K-1$ competing products are at the highest possible level for all $0 \le t \le \tau_k^K(\varepsilon)$, which slows down the convergence of the learning transient for product k (Theorem 3.1).

Furthermore, when prices are all set equal to p, using the fact that $\psi(-1,-\varepsilon)\sim 1/\varepsilon$ as $\varepsilon\to 0$, it can be seen that

$$\tau_k^K(\varepsilon)\sim \frac{B_{k,0}(\mathrm{e}^{p-1}+K)}{\varepsilon} \quad \text{as} \quad \varepsilon\to 0. \tag{B.4}$$

That is, for sufficiently small $\varepsilon>0$, the worst-case $\tau_k^K(\varepsilon)$ is roughly linear in the number of products K, consistently with the intuition discussed in Section 3.

B.2. Worst-Case Regret with Different Prices

Here we consider the more general setting where prices may be different. In this case, recall from (4.7) that the product ranking at any given time is determined by an index that depends both on the quality and price of each product. As such, the (worst-case) regret of the greedy policy does not admit a closed form. For the purpose of characterizing a lower bound of the worst-case regret, and in light of Proposition 4.2, it is reasonable to consider a product configuration where prior beliefs are $\hat{q}_{k,0}=q_k$ for all $k\neq 1$ and $\hat{q}_{1,0}=0$. The regret in this configuration would serve as a lower bound for the worst-case regret, which is characterized in the following theorem.

THEOREM B.1. *Consider the fluid model approximation in (3.1). Assume $B_{k,0}=B_0$ for each $k=1,\dots,K$. Consider the product configuration $Q^K\in\mathcal{Q}^K$ such that $\hat{q}_{k,0}=q_k$ for all $k\neq 1$ and $\hat{q}_{1,0}=0$. Let $q_{K+1}:=0$. Without loss of generality, assume $z_{k,\infty}=k$ for all $k=1,\dots,K$. Then, as $K\to\infty$,*

$$\mathfrak{R}_{\mathsf{G}}(Q^K)\sim\sum_{j=2}^{K}\sum_{k=2}^{j}((p_1-r_\infty)v_1-(p_k-r_\infty)v_k)\left(\mathrm{e}^{g(j)-g(k-1)}-\mathrm{e}^{g(j)-g(k)}\right)\psi(q_{j+1}-q_1,q_j-q_1). \tag{B.5}$$

Proof of Theorem B.1. Without loss of generality, we assume that $1=q_1\geq q_2\geq\cdots\geq q_K$. Define $s_j:=\inf\{t\geq 0: z_1(t)=j-1\}$ for each $j=2,\dots,K$ with $s_{K+1}:=0$; that is, s_j is the first time when the perceived quality of product 1 is the $(j-1)$-st highest. For the product configuration Q^K defined in the statement of the theorem, it is trivial to check that $0\equiv s_{K+1}\leq s_K\leq\cdots\leq s_2$. In the proof, we consider a sufficiently large $T\geq s_2$. To simplify the exposition, we define

$$\phi_k:=\begin{cases}\psi(q_{k+1}-q_1,q_k-q_1) & \text{for } k=2,\dots,K-1;\\ \psi(-q_1,q_k-q_1) & \text{for } k=K.\end{cases} \tag{B.6}$$

The proof of the theorem will be done in three steps. In the first step, we characterize the revenue under the greedy policy. In the second step, we characterize the revenue under the optimal policy. Finally, we derive the expression for the regret and the revenue gap between the greedy and optimal policies.

Step 1. We characterize the revenue under the greedy policy. Observe that

$$R_T^{\mathsf{G}}=\sum_{j=1}^{K}p_j\int_0^T d_j(\hat{\boldsymbol{q}}(t),\boldsymbol{z}(t))\,\mathrm{d}t=p_1\underbrace{\int_0^T d_1(\hat{\boldsymbol{q}}(t),\boldsymbol{z}(t))\,\mathrm{d}t}_{A_1}+\sum_{j=2}^{K}p_j\underbrace{\int_0^T d_j(\hat{\boldsymbol{q}}(t),\boldsymbol{z}(t))\,\mathrm{d}t}_{A_j}. \tag{B.7}$$

The first integral on the right-hand side of the preceding equation can be written as

$$\begin{aligned} A_1 &= \int_0^{s_2} d_1(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t + \int_{s_2}^{T} d_1(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t \\ &= B_0 \frac{q_2 - \hat{q}_1(0)}{q_1 - q_2} + \int_{s_2}^{T} d_1(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t, \end{aligned} \tag{B.8}$$

where the second equality follows from the definition of s_2, so that

$$\hat{q}_1(s_2) = q_2 = q_1 + (\hat{q}_1(0) - q_1)\frac{B_0}{B_1(s_2) + B_0} \iff B_1(s_2) = B_0 \frac{q_2 - \hat{q}_1(0)}{q_1 - q_2}. \tag{B.9}$$

Furthermore, the term A_j in (B.7) for $j \neq 1$ can be written as

$$A_j = \underbrace{\int_0^{s_j} d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t}_{A_j'} + \underbrace{\int_{s_j}^{s_2} d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t}_{A_j''} + \underbrace{\int_{s_2}^{T} d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t}_{A_j'''}. \tag{B.10}$$

We now derive expression for the individual terms in (B.10). First, A_j' can be written as

$$\begin{aligned} A_j' &= \mathrm{e}^{q_j - p_j - g(j-1)} \int_0^{s_j} \frac{\mathrm{d}t}{1 + \sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t) - p_k - z_k(t)}} = \mathrm{e}^{q_j - p_j - g(j-1)} \int_0^{s_j} \dot{B}_1(t) \mathrm{e}^{-\hat{q}_1(t) + p_1 + g(z_1(t))}\,\mathrm{d}t \\ &= \mathrm{e}^{q_j - p_j - g(j-1)} \sum_{k=j}^{K} \int_{s_{k+1}}^{s_k} \dot{B}_1(t) \mathrm{e}^{-\hat{q}_1(t) + p_1 + g(k)}\,\mathrm{d}t = \frac{v_j}{v_1} B_0 \sum_{k=j}^{K} \mathrm{e}^{g(k) - g(j-1)} \phi_k, \end{aligned} \tag{B.11}$$

where the last equation follows from Lemma C.1 and the definition $v_j \coloneqq \mathrm{e}^{q_j - p_j}$. Using the similar logical steps, the second integral A_j'' in (B.10) can be written as

$$\begin{aligned} A_j'' &= \mathrm{e}^{q_j - p_j - g(j)} \int_{s_j}^{s_2} \frac{1}{1 + \sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t) - p_k - z_k(t)}}\,\mathrm{d}t = \mathrm{e}^{q_j - p_j - g(j)} \int_{s_j}^{s_2} \dot{B}_1(t) \mathrm{e}^{-\hat{q}_1(t) + p_1 + g(z_1(t))}\,\mathrm{d}t \\ &= \mathrm{e}^{q_j - p_j - g(j)} \sum_{k=2}^{j-1} \int_{s_{k+1}}^{s_k} \dot{B}_1(t) \mathrm{e}^{-\hat{q}_1(t) + p_1 + g(k)}\,\mathrm{d}t = \frac{v_j}{v_1} B_0 \sum_{k=2}^{j-1} \mathrm{e}^{g(k) - g(j)} \phi_k. \end{aligned} \tag{B.12}$$

In the third integral A_j''' of (B.10), note that $\boldsymbol{z}(t) = \boldsymbol{z}_\infty$ for $t \geq s_2$. Combining these observations, we obtain

$$\frac{R_T^{\mathsf{G}}}{B_0} = p_1 \frac{q_2}{q_1 - q_2} + \sum_{j=2}^{K} p_j \frac{v_j}{v_1} \begin{pmatrix} \sum_{k=j}^{K} \mathrm{e}^{g(k) - g(j-1)} \phi_k \\ + \sum_{k=2}^{j-1} \mathrm{e}^{g(k) - g(j)} \phi_k \end{pmatrix} + \sum_{j=1}^{K} p_j \int_{s_2}^{T} d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}_\infty)\,\mathrm{d}t. \tag{B.13}$$

To derive an expression for τ, observe that

$$\begin{aligned} B_1(s_2) = B_0 \frac{q_2}{1 - q_2} &= \int_0^{s_2} \frac{\mathrm{e}^{\hat{q}_1(t) - p_1 - g(z_1(t)}}{1 + \sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t) - p_k - g(z_k(t))}}\,\mathrm{d}t \\ &= s_2 - \underbrace{\int_0^{s_2} \frac{1 + \sum_{j=2}^{K} \mathrm{e}^{q_j - p_j - g(z_j(t)}}{1 + \sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t) - p_k - g(z_k(t))}}\,\mathrm{d}t}_{\star}. \end{aligned} \tag{B.14}$$

Observe that

$$\begin{aligned}\star &= \sum_{i=2}^{K} \int_{s_{i+1}}^{s_i} \frac{1+\sum_{j=2}^{K} \mathrm{e}^{q_j-p_j-g(z_j(t))}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t)-p_k-g(z_k(t))}}\,\mathrm{d}t \\ &= \int_{s_{K+1}}^{s_K} \frac{1+\sum_{j=2}^{K} \mathrm{e}^{q_j-p_j-g(j-1)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t)-p_k-g(z_k(t))}}\,\mathrm{d}t + \sum_{i=2}^{K-1} \int_{s_{i+1}}^{s_i} \frac{1+\sum_{j=i+1}^{K} \mathrm{e}^{q_j-p_j-g(j)} + \sum_{j=2}^{i} \mathrm{e}^{q_j-p_j-g(j-1)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t)-p_k-g(k-1)}}\,\mathrm{d}t.\end{aligned} \tag{B.15}$$

The first integral in (B.15) can be written as

$$\begin{aligned}\int_{s_{K+1}}^{s_K} \frac{1+\sum_{j=2}^{K} \mathrm{e}^{q_j-p_j-g(j-1)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t)-p_k-g(z_k(t))}}\,\mathrm{d}t &= \left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right) \int_{s_{K+1}}^{s_K} \mathrm{e}^{p_1-\hat{q}_1(t)} \dot{B}_1(t)\,\mathrm{d}t \\ &= \frac{B_0}{v_1}\left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right)\phi_K,\end{aligned} \tag{B.16}$$

where the first equality follows from the definition $v_j = \mathrm{e}^{q_j-p_j}$ and the second follows from Lemma C.1. Following the similar steps, the second integral in (B.15) can be written as

$$\begin{aligned}&\sum_{i=2}^{K-1} \int_{s_{i+1}}^{s_i} \frac{1+\sum_{j=i+1}^{K} \mathrm{e}^{q_j-p_j-g(j)} + \sum_{j=2}^{i} \mathrm{e}^{q_j-p_j-g(j-1)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k(t)-p_k-g(k-1)}}\,\mathrm{d}t \\ &= \sum_{i=2}^{K-1} \begin{pmatrix} 1+\sum_{j=i+1}^{K} v_j \mathrm{e}^{g(i)-g(j)} \\ +\sum_{j=2}^{i} v_j \mathrm{e}^{g(i)-g(j-1)} \end{pmatrix} \int_{s_{i+1}}^{s_i} \mathrm{e}^{p_1-\hat{q}_1(t)} \dot{B}_1(t)\,\mathrm{d}t \\ &= \frac{B_0}{v_1} \sum_{i=2}^{K-1} \begin{pmatrix} 1+\sum_{j=i+1}^{K} v_j \mathrm{e}^{g(i)-g(j)} \\ +\sum_{j=2}^{i} v_j \mathrm{e}^{g(i)-g(j-1)} \end{pmatrix} \phi_i.\end{aligned} \tag{B.17}$$

Therefore, we deduce that

$$\begin{aligned}\frac{s_2}{B_0} = \frac{q_2}{q_1-q_2} &+ \frac{1}{v_1}\left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right)\phi_K \\ &+ \frac{1}{v_1} \sum_{i=2}^{K-1} \begin{pmatrix} 1+\sum_{j=i+1}^{K} v_j \mathrm{e}^{g(i)-g(j)} \\ +\sum_{j=2}^{i} v_j \mathrm{e}^{g(i)-g(j-1)} \end{pmatrix} \phi_i.\end{aligned} \tag{B.18}$$

Step 2. Consider the optimal policy such that $\boldsymbol{z}^*(t) = \boldsymbol{z}_\infty = (1, 2, \ldots, K)$ for each $t \geq 0$. Define $\tau^* := \inf\{t \colon \hat{q}_1^*(t) \geq q_2\}$. Using similar steps as *Step 1*, one can derive the expression for the revenue under the optimal policy. Specifically, one can write

$$\frac{R_T^*}{B_0} = p_1 \frac{q_2}{q_1-q_2} + \sum_{k=2}^{K} \sum_{j=2}^{K} p_j \frac{v_j}{v_1} \mathrm{e}^{g(1)-g(j)} \phi_k + \sum_{j=1}^{K} p_j \int_{s_2^*}^{T} d_j(\hat{q}^*(t), z_\infty)\,\mathrm{d}t. \tag{B.19}$$

To derive the expression for s_2^*, observe that

$$\begin{aligned}B_1^*(s_2^*) = B_0 \frac{q_2}{q_1-q_2} &= \int_0^{s_2^*} \frac{\mathrm{e}^{\hat{q}_1^*(t)-p_1-g(1)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k^*(t)-p_k-g(k)}}\,\mathrm{d}t \\ &= s_2^* - \underbrace{\int_0^{s_2^*} \frac{1+\sum_{j=2}^{K} \mathrm{e}^{q_j-p_j-g(j)}}{1+\sum_{k=1}^{K} \mathrm{e}^{\hat{q}_k^*(t)-p_k-g(k)}}\,\mathrm{d}t}_{\star'},\end{aligned} \tag{B.20}$$

where

$$\begin{aligned}\star' &= \left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right)\int_0^{s_2^*} \exp(p_1-\hat{q}_1(t))\dot{B}_1(t)\,\mathrm{d}t \\ &= \frac{B_0}{v_1}\left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right)\sum_{k=2}^{K}\phi_k,\end{aligned} \tag{B.21}$$

and the last equality follows from Lemma C.1. Thus, we obtain

$$\frac{s_2^*}{B_0} = \frac{q_2}{q_1-q_2} + \frac{1}{v_1}\left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right)\sum_{k=2}^{K}\phi_k. \tag{B.22}$$

Step 3. We characterize the difference in revenue between the greedy and optimal policies, characterized in (B.13) and (B.19), respectively. First, it is trivial to check that $s_2^* \le s_2$ because $z_1^*(t)=1$ for all t under the optimal policy while $z_1(t)>1$ until $t<s_2$ under the greedy policy. Also, observe that $d_j(\hat{q}(t),\boldsymbol{z}_\infty) = d_j(\hat{q}^*(t-s_2+s_2^*),\boldsymbol{z}_\infty)$ for all $t \ge s_2$. That is, the revenue under the greedy policy during $[s_2, s_2+s]$ is identical to the revenue under the optimal policy during $[s_2^*, s_2^*+s]$ for any $s \le T-s_2$. Therefore, the difference between the last terms in (B.13) and (B.19) can be written as

$$\begin{aligned}&\sum_{j=1}^{K} p_j \int_{s_2^*}^{T} d_j(\hat{\boldsymbol{q}}^*(t),\boldsymbol{z}_\infty)\,\mathrm{d}t - \sum_{j=1}^{K} p_j \int_{s_2}^{T} d_j(\hat{\boldsymbol{q}}(t),\boldsymbol{z}_\infty)\,\mathrm{d}t \\ &= \sum_{j=1}^{K} p_j \int_{T-s_2+s_2^*}^{T} d_j(\hat{\boldsymbol{q}}^*(t),\boldsymbol{z}_\infty)\,\mathrm{d}t \\ &= \tilde{r}_\infty(s_2-s_2^*),\end{aligned} \tag{B.23}$$

where $\tilde{r}_\infty \sim r_\infty$ as $K\to\infty$ and r_∞ is defined in (4.2). (To see this, note from (B.18) that $s_2\to\infty$ as $K\to\infty$, which, by construction, means $T\ge s_2$ also grows large as K increases.) To bound $s_2-s_2^*$, observe from (B.18) and (B.22) that

$$\begin{aligned}\frac{s_2-s_2^*}{B_0/v_1} = &\left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right)\phi_K + \sum_{i=2}^{K-1}\left(\begin{array}{c}1+\sum_{j=i+1}^{K} v_j \mathrm{e}^{g(K)-g(j)} \\ +\sum_{j=2}^{i} v_j \mathrm{e}^{g(K)-g(j-1)}\end{array}\right)\phi_n \\ &- \left(1+\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right)\sum_{k=2}^{K}\phi_k.\end{aligned} \tag{B.24}$$

Recall the definition $\mathfrak{R}_{\mathsf{G}} = \lim_{T\to\infty}\{R_T^* - R_T^{\mathsf{G}}\}$. Combining these into (B.13) and (B.19), we have, as $K\to\infty$,

$$\begin{aligned}\frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} \sim &\sum_{k=2}^{K}\sum_{j=2}^{K} p_j v_j \mathrm{e}^{g(1)-g(j)}\phi_k - \sum_{j=2}^{K} p_j v_j \left(\begin{array}{c}\sum_{k=j}^{K} \mathrm{e}^{g(k)-g(j-1)}\phi_k \\ +\sum_{k=2}^{j-1} \mathrm{e}^{g(k)-g(j)}\phi_k\end{array}\right) \\ &+ r_\infty \left(\begin{array}{c}\left(\mathrm{e}^{g(K)}+\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right)\phi_K \\ +\sum_{k=2}^{K-1}\left(\mathrm{e}^{g(k)}+\sum_{j=k+1}^{K} v_j \mathrm{e}^{g(k)-g(j)}+\sum_{j=2}^{k} v_j \mathrm{e}^{g(k)-g(j-1)}\right)\phi_k \\ -\left(\mathrm{e}^{g(1)}+\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right)\sum_{k=2}^{K}\phi_k\end{array}\right).\end{aligned} \tag{B.25}$$

By changing the order of summation, one can write

$$\sum_{j=2}^{K} p_j v_j \left(\begin{array}{l} \sum_{k=j}^{K} \mathrm{e}^{g(k)-g(j-1)}\phi_k \\ +\sum_{k=2}^{j-1} \mathrm{e}^{g(k)-g(j)}\phi_n \end{array} \right) = \sum_{k=2}^{K}\sum_{j=2}^{k} p_j v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k + \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} p_j v_j \mathrm{e}^{g(k)-g(j)}\phi_k. \tag{B.26}$$

Plugging this into (B.25) and recalling that $g(1)=0$, we deduce

$$\begin{aligned} \frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} \sim & \sum_{k=2}^{K}\sum_{j=2}^{K} p_j v_j \mathrm{e}^{-g(j)}\phi_k - \sum_{k=2}^{K}\sum_{j=2}^{k} p_j v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k - \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} p_j v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\ & + r_\infty \left(\begin{array}{c} \left(\mathrm{e}^{g(K)} + \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right)\phi_K \\ + \sum_{k=2}^{K-1} \left(\mathrm{e}^{g(k)} + \sum_{j=k+1}^{K} v_j \mathrm{e}^{g(k)-g(j)} + \sum_{j=2}^{k} v_j \mathrm{e}^{g(k)-g(j-1)}\right)\phi_k \\ - \left(1 + \sum_{j=2}^{K} v_j \mathrm{e}^{-g(j)}\right)\sum_{k=2}^{K}\phi_k \end{array} \right). \end{aligned} \tag{B.27}$$

Rearranging terms, one can write

$$\begin{aligned} \frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} \sim & \sum_{k=2}^{K}\sum_{j=2}^{K} p_j v_j \mathrm{e}^{-g(j)}\phi_k - \sum_{k=2}^{K}\sum_{j=2}^{n} p_j v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k - \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} p_j v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\ & + r_\infty \sum_{k=2}^{K} \mathrm{e}^{g(k)}\phi_k + r_\infty \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} v_j \mathrm{e}^{g(k)-g(j)}\phi_n + r_\infty \sum_{k=2}^{K}\sum_{j=2}^{k} v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k \\ & - r_\infty \sum_{k=2}^{K}\phi_k - r_\infty \sum_{k=2}^{K}\sum_{j=2}^{K} v_j \mathrm{e}^{-g(j)}\phi_k. \end{aligned} \tag{B.28}$$

Collecting terms using factors $\mathrm{e}^{g(k)-g(j)}$ and $\mathrm{e}^{g(k)-g(j-1)}$, as $K\to\infty$, the above can be reformulated as

$$\begin{aligned} \frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} \sim & \, r_\infty \sum_{k=2}^{K}\sum_{j=2}^{k} v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k - \sum_{k=2}^{K}\sum_{j=2}^{k} p_j v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k \\ & + r_\infty \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} v_j \mathrm{e}^{g(k)-g(j)}\phi_k - \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} p_j v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\ & - r_\infty \sum_{k=2}^{K}\sum_{j=2}^{K} v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}\sum_{j=2}^{K} p_j v_j \mathrm{e}^{-g(j)}\phi_k + r_\infty \sum_{k=2}^{K} \mathrm{e}^{g(k)}\phi_k - r_\infty \sum_{k=2}^{K}\phi_k \\ = & \sum_{k=2}^{K}\sum_{j=2}^{k} (r_\infty - p_j) v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k + \sum_{k=2}^{K-1}\sum_{j=k+1}^{K} (r_\infty - p_j) v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\ & - \sum_{k=2}^{K}\sum_{j=2}^{K} (r_\infty - p_j) v_j \mathrm{e}^{-g(j)}\phi_k + r_\infty \sum_{k=2}^{K} \mathrm{e}^{g(k)}\phi_k - r_\infty \sum_{k=2}^{K}\phi_k. \end{aligned} \tag{B.29}$$

Using the relation $r_\infty = \sum_{j=1}^{K}(p_j - r_\infty)v_j \mathrm{e}^{-g(j)}$, the above reduces to

$$\begin{aligned}
&\sum_{k=2}^{K}\sum_{j=2}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k + \sum_{k=2}^{K-1}\sum_{j=k+1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\
&- \sum_{k=2}^{K}\sum_{j=2}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}\sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k \\
&- \sum_{k=2}^{K}\sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k.
\end{aligned} \tag{B.30}$$

Rewriting the last two summations, we obtain

$$\begin{aligned}
&\sum_{k=2}^{K}\sum_{j=2}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k + \sum_{k=2}^{K-1}\sum_{j=k+1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\
&- \sum_{k=2}^{K}\sum_{j=2}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}\sum_{j=2}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}(r_\infty - p_1)v_1\phi_k \\
&- \sum_{k=2}^{K-1}\sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k - \sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(K)-g(j)}\phi_K.
\end{aligned} \tag{B.31}$$

Likewise, decomposing the first summation, the preceding expression can be written as

$$\begin{aligned}
&\sum_{k=2}^{K-1}\sum_{j=2}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k + \sum_{j=2}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(K)-g(j-1)}\phi_K \\
&- \sum_{k=2}^{K}\sum_{j=2}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}\sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{-g(j)}\phi_k + \sum_{k=2}^{K}(r_\infty - p_1)v_1\phi_k \\
&- \sum_{k=2}^{K-1}\sum_{j=1}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k - \sum_{j=1}^{K}(r_\infty - p_j)v_j \mathrm{e}^{g(K)-g(j)}\phi_K.
\end{aligned} \tag{B.32}$$

Rearranging terms, the preceding expression further reduces to

$$\begin{aligned}
&\sum_{k=2}^{K}\sum_{j=2}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j-1)}\phi_k - \sum_{k=2}^{K}\sum_{j=2}^{k}(r_\infty - p_j)v_j \mathrm{e}^{g(k)-g(j)}\phi_k \\
&+ \sum_{k=2}^{K}(r_\infty - p_1)v_1\phi_k - \sum_{k=2}^{K-1}(r_\infty - p_1)v_1 \mathrm{e}^{g(k)-g(1)}\phi_k.
\end{aligned} \tag{B.33}$$

Thus, we establish that, as $K \to \infty$,

$$\begin{aligned}
\frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} &\sim \sum_{k=2}^{K}(r_\infty - p_1)v_1(1 - \mathrm{e}^{g(k)})\phi_k - \sum_{k=2}^{K}\sum_{j=2}^{k}(r_\infty - p_j)v_j(\mathrm{e}^{g(k)-g(j)} - \mathrm{e}^{g(k)-g(j-1)})\phi_k \\
&= \sum_{k=2}^{K}\sum_{j=2}^{k}\left((p_1 - r_\infty)v_1 - (p_j - r_\infty)v_j\right)\left(\mathrm{e}^{g(k)-g(j-1)} - \mathrm{e}^{g(k)-g(j)}\right)\phi_k,
\end{aligned} \tag{B.34}$$

where the last equality follows from the fact that $\sum_{j=2}^{k}(\mathrm{e}^{g(k)-g(j-1)} - \mathrm{e}^{g(k)-g(j)}) = \mathrm{e}^{g(k)} - 1$. This completes the proof of the theorem. $\square$

An immediate corollary of Theorem B.1 is that the $\mathfrak{R}_{\mathsf{G}}(Q^K)$ is proportional to $\mathrm{e}^{g(K)}$. To see this, observe that

$$\begin{aligned}
&\sum_{j=2}^{K}\sum_{k=2}^{j}\left((p_1-r_\infty)v_1-(p_k-r_\infty)v_k\right)\left(\mathrm{e}^{g(j)-g(k-1)}-\mathrm{e}^{g(j)-g(k)}\right)\psi(q_{j+1}-1,q_j-1)\\
\geq&\left((p_1-r_\infty)v_1-(p_2-r_\infty)v_2\right)\sum_{j=2}^{K}\mathrm{e}^{g(j)}\psi(q_{j+1}-q_1,q_j-q_1)\sum_{k=2}^{j}\left(\mathrm{e}^{-g(k-1)}-\mathrm{e}^{-g(k)}\right)\\
\geq&\left((p_1-r_\infty)v_1-(p_2-r_\infty)v_2\right)\sum_{j=2}^{K}\psi(q_{j+1}-q_1,q_j-q_1)\left(\mathrm{e}^{g(j)-g(1)}-1\right)\\
\geq&\left((p_1-r_\infty)v_1-(p_2-r_\infty)v_2\right)\psi(-q_1,q_2-q_1)\left(\mathrm{e}^{g(K)-g(1)}-1\right)
\end{aligned}\tag{B.35}$$

where the first inequality follows from the fact that $(p_1-r_\infty)v_1\geq(p_2-r_\infty)v_2\geq\cdots\geq(p_K-r_\infty)v_K$ from the characterization of z_∞ in (4.7). Since $\mathfrak{R}_{\mathsf{G}}(Q^K)$ serves as a lower bound for the worst-case regret, we deduce that the worst-case regret $\mathfrak{R}_{\mathsf{G}}^K$ is at least of order $\mathrm{e}^{g(K)}$. Note that these observations are consistent with Theorem 4.1 in the case with equal prices.

Appendix C: Proofs for the Main Theoretical Results

In the proofs, we define $v_k:=\mathrm{e}^{q_k-p_k}$ for each $k=1,\dots,K$ to simplify notation.

C.1. Proofs for Section 3

Proof of Theorem 3.1. In the proof, we fix $\varepsilon>0$ and omit in function arguments to improve clarity of exposition. Recall from (3.3) that $\hat{q}_k(t)=q_k-(q_k-\hat{q}_{k,0})B_{k,0}/(B_{k,0}+B_k(t))$. By Assumption 2.1, we have that $B_k(t)\to\infty$ as $t\to\infty$, and therefore, it follows that $\hat{q}_k(t)\to q_k$ for each k as $t\to\infty$. We will next prove that τ_k^K is strictly decreasing in $\hat{q}_{k,0}$ for $\hat{q}_{k,0}<q_k$. To this end, consider the following control problem:

$$\begin{aligned}
\max_{u(t)\in[\underline{q}_0,\bar{q}_0]}\quad&\hat{q}_k(T)\\
\text{subject to}\quad&\dot{B}_j(t)=d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)))\quad\text{for each}\quad j=1,\dots,K\\
&\hat{Q}_k(B_k(t))=q_k-(q_k-u(t))\frac{B_{k,0}}{B_{k,0}+B_k(t)},
\end{aligned}\tag{C.1}$$

where $\underline{q}_0$ and $\bar{q}_0$ are arbitrary constants such that $\underline{q}_0<\bar{q}_0<q_k$. The desired result would follow if we show that the static policy $u^*(t)=\bar{q}_0$, $t\in[0,T]$, is optimal for an *arbitrary* $T>0$. Notice that a solution to (C.1) exists because the objective function, being independent of u, is trivially a concave function of the control variables, and because the control space $[\underline{q}_0,\bar{q}_0]$ is a compact set. This shows that the conditions of Theorem 1 in Cesari (1966) are satisfied, and that a solution to (C.1) exists. To characterize the solution of (C.1), define the Hamiltonian function $H(\boldsymbol{B}(t),\boldsymbol{\mu}(t),u(t)):=\sum_{k=1}^K\mu_k(t)d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)))$, where the costate vector $\boldsymbol{\mu}(t)$ satisfies the transversality condition; that is, $\mu_j(T)=0$ for $j\neq k$ and

$$\mu_k(T)=\dot{Q}_k(B_k^*(T))=(q_k-u^*(T))\frac{B_{k,0}}{(B_{k,0}+B_k^*(T))^2}.\tag{C.2}$$

Note that the Hamiltonian is constant over time under the optimal solution, so we define a constant $h \coloneqq H(\boldsymbol{B}^*(t), \boldsymbol{\mu}(t), u^*(t))$, $t \in [0,T]$. This implies that for $t = T$, $h = \mu_k(T) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}^*(T)))$ such that $\mu_k(T) - h = \frac{h}{d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}^*(T)))} - h > 0$. Furthermore, according to the Pontrygin Maximum Principle (PMP), the costate variable must satisfy, for each $j = 1, \ldots, K$,

$$\begin{aligned} \dot{\mu}_j(t) &= -\frac{\partial}{\partial B_j} H(\boldsymbol{B}^*(t), \boldsymbol{\mu}(t), u^*(t)) \\ &= -\dot{\hat{\boldsymbol{Q}}}(B_j^*(t))(\mu_j(t) - h). \end{aligned} \tag{C.3}$$

Note that $\mu_j(t)$ cannot cross h; specifically, when $\mu_j(t) = h$ for some t, then $\dot{\mu}_j(t) = 0$ and hence the costate variable is fixed to h till the end of the horizon, violating the transversaility condition. Recalling the fact that $\mu_k(T) - h > 0$, we have that $\mu_k(t) - h > 0$ for all $t \in [0,T]$. Moreover, from PMP, the optimal solution must satisfy that $u^*(t) = \arg\max_{v \in [\underline{q}, \bar{q}]} \{H(\boldsymbol{B}^*(t), \boldsymbol{\mu}^*(t), v)\}$. From the fact that $\mu_k(t) - h > 0$ for $t \in [0,T]$, we deduce

$$\frac{\partial}{\partial v} H(\boldsymbol{B}(t), \boldsymbol{\mu}(t), v)\Big|_{v=u(t)} = \frac{B_{k,0}}{B_{k,0} + B_k(t)} d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)))(\mu_k(t) - h) > 0. \tag{C.4}$$

This implies that $u^*(t) = \bar{q}_0$ for each $t \in [0,T]$ and the desired result follows. To show that τ_k^K is strictly increasing in $\hat{q}_{k,0}$ for $\hat{q}_{k,0} > q_k$, one may consider the minimization (instead of maximization) in (C.1). Moreover, the above proof strategy can be easily adapted to prove the strict monotonicity of τ_k^K with respect to q_k (considering the cases $q_k \geq \hat{q}_{k,0}$ and $q_k < \hat{q}_{k,0}$ separately), p_k, $B_{k,0}$, q_j, $\hat{q}_{j,0}$, p_j, and $B_{j,0}$ for $j \neq k$. We thus omit the details of these proofs for the sake of space.

It remains to prove that τ_k^K is strictly increasing in K. To this end, consider two nested markets S^K and S^{K+1} and let τ_k^K and τ_k^{K+1} be respectively the time-to-learn for product k in S_K and S_{K+1} respectively. Now, notice that S_K is equivalent to a $(K+1)$-dimensional market $\tilde{S}_{K+1}$ where the price $\tilde{p}_{K+1}$ of the $(K+1)$-th product is set equal to $+\infty$. Suppose now that $k \leq K$ and let $\tilde{\tau}_k^{K+1}$ be the time-to-learn for product $k \leq K$ in $\tilde{S}_{K+1}$; notice that $\tau_k^K = \tilde{\tau}_k^{K+1}$. The desired result then follows because, as we proved in the first part of the proof, $\tilde{\tau}_k$ is strictly decreasing in p_{K+1} and we clearly have $\tau_k^K = \tilde{\tau}_k^{K+1} \leq \tau_k^{K+1}$. This proves that the time-to-learn for product k is strictly decreasing in K, and concludes the proof of Theorem 3.1. □

The following lemma will be useful for the proofs of our main theoretical results. For the proofs for Section 3, one may apply the following lemma with zero search cost; that is, $g(z) \coloneqq 0$ for all $z \geq 1$.

LEMMA C.1. *Let $\hat{\boldsymbol{q}}(t) = (\hat{q}_1(t), \ldots, \hat{q}_K(t))$ be the vector of quality estimates in the fluid approximation in (3.3). Then, for any vector $\boldsymbol{z} = (z_1, \ldots, z_K)$, $0 \leq t_1 \leq t_2$, and any $k = 1, \ldots, K$, we have*

$$\int_{t_1}^{t_2} \frac{\mathrm{d}t}{1 + \sum_{j=1}^K \mathrm{e}^{\hat{q}_j(t) - p_j - g(z_j)}} = B_{k,0}(q_k - \hat{q}_{k,0}) \mathrm{e}^{p_k + g(z_k) - q_k} \psi(\hat{q}_k(t_1) - q_k, \hat{q}_k(t_2) - q_k), \tag{C.5}$$

where the function $\psi(\cdot)$ has been defined in (4.12).

Proof of Lemma C.1. Using the definition of the demand function in (3.3), we obtain

$$\int_{t_1}^{t_2} \frac{\mathrm{d}t}{1+\sum_{j=1}^{K} \mathrm{e}^{\hat{q}_j(t)-p_j-g(z_j)}} = \mathrm{e}^{p_k+g(z_k)} \int_{t_1}^{t_2} \mathrm{e}^{-\hat{q}_k(t)} \dot{B}_k(t)\,\mathrm{d}t. \tag{C.6}$$

The above equality holds for all $k=1,\dots,K$. Fix any k and notice that we can write

$$\begin{aligned}
\mathrm{e}^{p_k+g(z_k)} \int_{t_1}^{t_2} \mathrm{e}^{-\hat{q}_k(t)} \dot{B}_k(t)\,\mathrm{d}t &= \mathrm{e}^{p_k+g(z_k)} \int_{t_1}^{t_2} \mathrm{e}^{q_k-(q_k-\hat{q}_{k,0})\frac{B_{k,0}}{B_{k,0}+B_k(t)}} \dot{B}_k(t)\,\mathrm{d}t \\
&= \mathrm{e}^{p_k+g(z_k)} \int_{B_k(t_1)}^{B_k(t_2)} \mathrm{e}^{q_k-(q_k-\hat{q}_{k,0})\frac{B_{k,0}}{B_{k,0}+x}}\,\mathrm{d}x \\
&= B_{k,0}(q_k-\hat{q}_{k,0})\mathrm{e}^{p_k+g(z_k)-q_k} \int_{\hat{q}_k(t_1)-q_k}^{\hat{q}_k(t_2)-q_k} \frac{\mathrm{e}^{-y}}{y^2}\,\mathrm{d}y,
\end{aligned} \tag{C.7}$$

where we used the substitution $y=(q_k-\hat{q}_{k,0})B_{k,0}/(B_{k,0}+x)$ in the second equation. This completes the proof of the lemma. □

C.2. Proofs for Section 4

C.2.1. Proofs for Section 4.2.

Proof of Proposition 4.1. First, we reformulate the platform's optimal control problem, rewriting it in terms of the vector of purchases $\boldsymbol{B}(t)$. Define $\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)) \coloneqq (\hat{Q}_1(\boldsymbol{B}(t)),\dots,\hat{Q}_K(\boldsymbol{B}(t))$, where

$$\hat{Q}_k(\boldsymbol{B}(t)) \coloneqq q_k - (q_k-\hat{q}_{k,0})\frac{B_{k,0}}{B_{k,0}+B_k(t)}. \tag{C.8}$$

Then, we can see that solving (4.1) is equivalent to solving the following optimal control problem

$$\begin{aligned}
\underset{\{\Pi(t)\}}{\text{maximize}} \quad & \sum_{k=1}^{K} p_k \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}^K} \int_0^T \pi_z(t) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)),\boldsymbol{z})\,\mathrm{d}t \\
\text{subject to} \quad & \dot{B}_k(t) = \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_z(t) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)),\boldsymbol{z}), \quad k=1,\dots,K \\
& \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) = 1, \quad t\in[0,T].
\end{aligned} \tag{C.9}$$

Note that the quality estimate vector $\hat{\boldsymbol{q}}(t)$ no longer appears in the above formulation of the platform's optimal control problem.

Having said that, we now prove that an optimal solution to (C.9) exists. Observe that the reachable set at any time t is bounded as $B_k(t)=\int_0^t \dot{B}_k(s)\,\mathrm{d}s < t$ for all k, and that the set of admissible velocities

$$V(t,\boldsymbol{B}) = \left\{(\dot{B}_1(t),\dot{B}_2(t),\dots,\dot{B}_K(t)) : 0\le\pi_k(t)\le 1,\ \forall k\in\mathcal{P}_K, \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K}\pi_{\boldsymbol{z}}(t)=1\right\},$$

is convex. The set $V(t,\boldsymbol{B})$ can indeed be seen as the set of all convex combinations of the vectors

$$\begin{pmatrix} d_1(\hat{Q}_1(\boldsymbol{B(t)},\boldsymbol{z}_1)) \\ d_2(\hat{Q}_2(\boldsymbol{B(t)},\boldsymbol{z}_1)) \\ \vdots \\ d_K(\hat{Q}_K(\boldsymbol{B(t)},\boldsymbol{z}_1)) \end{pmatrix}, \begin{pmatrix} d_1(\hat{Q}_1(\boldsymbol{B(t)},\boldsymbol{z}_2)) \\ d_2(\hat{Q}_2(\boldsymbol{B(t)},\boldsymbol{z}_2)) \\ \vdots \\ d_K(\hat{Q}_K(\boldsymbol{B(t)},\boldsymbol{z}_2)) \end{pmatrix}, \dots \begin{pmatrix} d_1(\hat{Q}_1(\boldsymbol{B(t)},\boldsymbol{z}_{K!})) \\ d_2(\hat{Q}_2(\boldsymbol{B(t)},\boldsymbol{z}_{K!})) \\ \vdots \\ d_K(\hat{Q}_K(\boldsymbol{B(t)},\boldsymbol{z}_{K!})) \end{pmatrix}.$$

Moreover, notice that $\dot{B}_k(t) \leq 1 + |\boldsymbol{B}(t)|$ holds trivially as $\dot{B}_k(t) = \sum_{\boldsymbol{z} \in \mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) d_k(\boldsymbol{B}(t), \boldsymbol{z})$ is a probability. In particular, this implies that the hypotheses of Theorem 5.1.1 of Bressan and Piccoli (2007) are satisfied, and that there exist an optimal solution for (C.9).

Now we proceed with the characterization of the optimal solution of (C.9). To do that, we define the vector of costate variables $\boldsymbol{\mu}(t) \coloneqq (\mu_1(t), \mu_2(t), \ldots, \mu_K(t))$, which satisfies the transversality condition $\mu_k(T) = 0$ for all $k \in \mathcal{P}_K$. The Hamiltonian function is defined as follows

$$H(\Pi(t), \boldsymbol{B}(t), \boldsymbol{\mu}(t)) \coloneqq \sum\nolimits_{\boldsymbol{z} \in \mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) \sum_{k=1}^{K} (p_k + \mu_k(t)) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}(t)), \boldsymbol{z}).$$

An optimal solution is denoted by $(\boldsymbol{B}^*(t), \Pi^*(t), \boldsymbol{\mu}^*(t))$. By the Pontryagin Maximum Principle (PMP), we know that any optimal solution satisfies the first order conditions

$$\Pi^*(t) = \arg\max\nolimits_{\{\Pi \in \Delta(\mathcal{Z}^K)\}} H(\Pi, \boldsymbol{B}^*(t), \boldsymbol{\mu}^*(t)) \tag{C.10}$$

$$\dot{\mu}_k^*(t) = -\frac{\partial}{\partial B_k} H(\Pi^*(t), \boldsymbol{B}^*(t), \boldsymbol{\mu}^*(t)) \tag{C.11}$$

$$\dot{B}_k^*(t) = \frac{\partial}{\partial \mu_k} H(\Pi^*(t), \boldsymbol{B}^*(t), \boldsymbol{\mu}^*(t)) \tag{C.12}$$

for all $k = 1, \ldots, K$. Moreover, the PMP guarantees that the Hamiltonian function, when evaluated in any optimal solution $(\boldsymbol{B}^*(t), \Pi^*(t), \boldsymbol{\mu}^*(t))$, is constant over time, i.e., there exists $h \in \mathbb{R}$ such that $H(\Pi^*(t), \boldsymbol{B}^*(t), \boldsymbol{\mu}^*(t)) = h$ for all $0 \leq t \leq T$. In the remainder of the section, for notational convenience, we omit the dependence from t. We now analyze the conditions (C.10), (C.11), and (C.12) separately.

Condition (C.10). We can use the result of Lemma C.2 to find that, for all $t \in [0, T]$, we have

$$\arg\max\nolimits_{\{\Pi \in \Delta(\mathcal{Z}^K)\}} H(\Pi, \boldsymbol{B}^*, \boldsymbol{\mu}^*) = \arg\max\nolimits_{\boldsymbol{z} \in \mathcal{Z}^K} \sum_{k=1}^{K} (p_k + \mu_k) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}), \boldsymbol{z}).$$

Let

$$\bar{\boldsymbol{Z}} = \arg\max\nolimits_{\boldsymbol{z} \in \mathcal{Z}^K} \sum_{k=1}^{K} (p_k + \mu_k) d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}), \boldsymbol{z}).$$

Then, it is easy to see that any probability distribution such that

$$\Pi^* = \begin{cases} \pi_{\boldsymbol{z}} \geq 0 & \text{if } \boldsymbol{z} \in \bar{\boldsymbol{Z}}, \\ 0 & \text{if } \boldsymbol{z} \notin \bar{\boldsymbol{Z}}, \end{cases} \tag{C.13}$$

where $\sum_{\boldsymbol{z} \in \bar{\boldsymbol{Z}}} \pi_{\boldsymbol{z}} = 1$, is a candidate solution of (C.10). Moreover, by the the transversality condition, evaluating the Hamiltonian in $t = T$ yields

$$\begin{aligned} h &= \max\nolimits_{\{\Pi\}} \sum\nolimits_{\boldsymbol{z} \in \mathcal{Z}_K} \pi_{\boldsymbol{z}} \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(T), \boldsymbol{z}) = \max\nolimits_{\boldsymbol{z} \in \mathcal{Z}_k} \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(T), \boldsymbol{z}) \\ &= \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(T), \boldsymbol{z}_\infty). \end{aligned} \tag{C.14}$$

Since the Hamiltonian is constant over the optimal path, from (C.14) we can conclude that

$$\sum_{k=1}^{K}(p_k+\mu_k)d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z}') = \max_{\boldsymbol{z}\in\mathcal{Z}^K}\sum_{k=1}^{K}(p_k+\mu_k)d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z}) = \sum_{k=1}^{K}p_k d_k(\hat{\boldsymbol{q}}(T),\boldsymbol{z}_\infty)) = h, \tag{C.15}$$

for any $\boldsymbol{z}'\in\bar{\boldsymbol{Z}}$.

Condition (C.11). For $j\neq k$ we have

$$\frac{\partial}{\partial B_j}d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z}) = -d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})\frac{\partial}{\partial B_j}\hat{Q}_j(B_j),$$
$$\frac{\partial}{\partial B_j}d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z}) = d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})(1-d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z}))\frac{\partial}{\partial B_j}\hat{Q}_j(B_j).$$

Then

$$\frac{\partial}{\partial B_j}H(\Pi,\boldsymbol{B},\boldsymbol{\mu}) = \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K}\pi_{\boldsymbol{z}}\sum_{k=1}^{K}(p_k+\mu_k)\frac{\partial}{\partial B_j}d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})$$
$$= \left\{\sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K}\pi_{\boldsymbol{z}}\Big[(p_j+\mu_j)-\sum_{k=1}^{K}(p_k+\mu_k)d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})\Big]\right\}d_j(\hat{\boldsymbol{Q}}(\boldsymbol{B}),\boldsymbol{z})\frac{\partial}{\partial B_j}\hat{Q}_j(B_j). \tag{C.16}$$

Plugging (C.13) and (C.15) into the above relation, we can see that, when evaluated over the optimal path, (C.11) can be reformulated as follows

$$\dot{\mu}_j^* = -\frac{\partial}{\partial B_j}H(\Pi^*,\boldsymbol{B}^*,\boldsymbol{\mu}^*) = -(p_j+\mu_j^*-h)\dot{B}_j^*\frac{\partial}{\partial B_j}\hat{Q}_j(B_j^*) = -(p_j+\mu_j^*-h)\dot{\hat{Q}}_j(B_j^*),$$

which, rewritten as $-\dot{\mu}_j^*/(p_j-h+\mu_j^*) = \dot{\hat{Q}}_j(B_j^*)$, can be integrated over $[0,T]$ on both sides to obtain

$$p_j+\mu_j^*(t) = h+(p_j-h)\exp\left[\hat{Q}_j(B_j^*(T))-\hat{Q}_j(B_j^*(t))\right],$$

where we restored the explicit dependence on time. The instantaneous optimal ranking $\bar{\boldsymbol{z}}(t)$ for $t\in[0,T]$ then satisfies

$$\bar{\boldsymbol{z}}(t) = \operatorname*{arg\,max}_{\boldsymbol{z}\in\mathcal{Z}_K}\sum_{k=1}^{K}\left\{h+(p_k-h)\exp\left[\hat{Q}_k(B_k^*(T))-\hat{Q}_k(B_k^*(t))\right]\right\}d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}^*(t)),\boldsymbol{z}). \tag{C.17}$$

The above combinatorial optimization problem can be seen as a MNLPP where the profit for product k at time t is $\rho_k(t) = h+(p_k-h)\exp\left[\hat{Q}_k(B_k^*(T))-\hat{Q}_k(B_k^*(t))\right]$. Using Theorem 1 in Abeliuk et al. (2016), we can show that the optimal any position assignment z^* and the corresponding optimal profit ρ^* must satisfy the following condition for all $t\in[0,T]$:

$$z_{k_1}^*(t) < z_{k_2}^*(t) \iff (\rho_{k_1}(t)-\rho^*)\exp(Q_{k_1}(B_{k_1}^*(t))) > (\rho_{k_2}(t)-\rho^*)\exp(Q_{k_2}(B_{k_2}^*(t))). \tag{C.18}$$

Notice that the optimal profit ρ^* for (C.17) is given by the value of the Hamiltonian function, which is constant when evaluated over the optimal solution, i.e.,

$$\rho^* = \max_{\boldsymbol{z}\in\mathcal{Z}_K} \sum_{k=1}^{K} \left\{ h + (p_k - h) \exp\left[\hat{Q}_k(B_k^*(T)) - \hat{Q}_k(B_k^*(t))\right] \right\} d_k(\hat{\boldsymbol{Q}}(\boldsymbol{B}^*(t)), \boldsymbol{z}) = h,$$

Hence, condition (C.18) translates into

$$\bar{z}_{k_1}(t) < \bar{z}_{k_2}(t) \iff (p_{k_1} - h) \exp\left[\hat{Q}_{k_1}(B_{k_1}^*(T)) - p_{k_1}\right] > (p_{k_2} - h) \exp\left[\hat{Q}_{k_2}(B_{k_2}^*(T)) - p_{k_2}\right],$$

Notice that the above condition implies that $\bar{\boldsymbol{z}}(t) = \boldsymbol{z}_\infty$ and that the r.h.s. does not depend on t. This implies that the above condition holds for all $t \in [0,T]$, which implies that the solution to the optimal control problem is static throughout the selling horizon and assigns maximum probability to the asymptotically optimal ranking $\boldsymbol{z}_\infty$. □

C.2.2. Proofs for Section 4.3. For the proofs in Section 4.3, we use the asymptotics with respect to the number of product K. Hence, most variables depend on K, but we suppress the dependence to improve clarity.

The following lemma will be used for the proofs in this section.

LEMMA C.2. *Consider the fluid model approximation in (3.1), and define Π^{G} as in (4.5). Then, (4.6) holds true for all $t \geq 0$.*

Proof of Lemma C.2. Recall that $\pi_{\boldsymbol{z}}(t) \coloneqq \mathsf{P}(\boldsymbol{\sigma}(t) = \boldsymbol{z})$. Then,

$$\mathsf{E}_{\Pi(t)}\left[\sum_{k=1}^{K} p_k \tilde{d}_k(\hat{\boldsymbol{q}}(t), \Pi(t))\right] = \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}).$$

Suppose that there exists $\bar{\boldsymbol{z}} \in \mathcal{Z}_K$ such that $\sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) \leq \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \bar{\boldsymbol{z}})$ for all $\boldsymbol{z} \in \mathcal{Z}_K$. Then, since $\sum_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) = 1$ for all t, we have

$$\sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) \leq \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \bar{\boldsymbol{z}}) \sum\nolimits_{\boldsymbol{z}\in\mathcal{Z}_K} \pi_{\boldsymbol{z}}(t) = \sum_{k=1}^{K} p_k d_k(\hat{\boldsymbol{q}}(t), \bar{\boldsymbol{z}}),$$

which gives the desired result and concludes the proof. □

LEMMA C.3. *Suppose that Assumption 4.1 holds. Fix a quality configuration $Q^K \in \mathcal{Q}^K$ and consider $k \neq 1$. Then, under the greedy policy, there exists $s > 0$ such that (i) $\hat{q}_1(s) = \hat{q}_k(s)$, (ii) $\hat{q}_1(t) < \hat{q}_k(t)$ for all $t < s$, and (iii) $\hat{q}_1(t) > \hat{q}_k(t)$ for all $t > s$.*

Proof of Lemma C.3. By Assumption 4.1, we have that $\hat{q}_1(0) < q_k < q_1$. If $q_k < \hat{q}_k(0)$, then the desired result immediately follows from the fact that $\hat{q}_k(t)$ and $\hat{q}_1(t)$ are monotonically decreasing and increasing, respectively. Therefore, it suffices to consider $q_k > \hat{q}_k(0)$. We consider two cases: (i) $\hat{q}_k(0) > \hat{q}_1(0)$ and (ii) $\hat{q}_k(0) \leq \hat{q}_1(0)$.

In case (i), toward a contradiction, suppose that there exist multiple crossing points. This implies that there exist $s_1 < s_2 < s_3$ such that $\hat{q}_1(s) = \hat{q}_k(s)$ for $s = s_1, s_2, s_3$. By construction, it follows that $\hat{q}_k(t) > \hat{q}_1(t)$ if and only if $t \in [0, s_1) \cup (s_2, s_3)$. By continuity of the paths, we have that $\dot{\hat{q}}_k(s_1) < \dot{\hat{q}}_1(s_1)$ and $\dot{\hat{q}}_k(s_2) > \dot{\hat{q}}_1(s_2)$. Let $\bar{q}_1 := \hat{q}_k(s_1) = \hat{q}_1(s_1)$ and $\bar{q}_2 := \hat{q}_k(s_2) = \hat{q}_1(s_k)$.

From (3.3), it can be seen that $\dot{\hat{q}}_j(t) = \frac{\dot{B}_j(t)(q_j - \hat{q}_j(t))^2}{B_0(q_j - \hat{q}_j(0))}$ for any $j = 1, \dots, K$. Also note that for $t < s_1$, $\dot{B}_k(t) > \dot{B}_1(t)$ because $\dot{B}_j(t) = d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))$ for any $j = 1, \dots, K$, where $d_j(\cdot)$ is defined in (2.3) and $\hat{q}_1(t) < \hat{q}_k(t)$ and $g(z_1(t)) > g(z_k(t))$ under the greedy policy. By the continuity of the path, we deduce that $\dot{B}_k(s_1) \geq \dot{B}_1(s_1)$. Combining these observations, we establish

$$\dot{\hat{q}}_1(t) > \dot{\hat{q}}_k(t) \Longrightarrow \frac{(q_1 - \bar{q}_1)^2}{(q_k - \bar{q}_1)^2} > \frac{q_1 - \hat{q}_1(0)}{q_k - \hat{q}_k(0)}. \tag{C.19}$$

Likewise, it can be seen that, for $t \in (s_1, s_2)$, $\dot{B}_k(t) < \dot{B}_1(t)$, because $\hat{q}_1(t) > \hat{q}_k(t)$ and $g(z_1(t)) < g(z_k(t))$ under the greedy policy. By continuity, we have $\dot{B}_k(s_2) \leq \dot{B}_1(s_2)$. Combining these observations, we deduce that

$$\dot{\hat{q}}_1(t) < \dot{\hat{q}}_k(t) \Longrightarrow \frac{(q_1 - \bar{q}_2)^2}{(q_k - \bar{q}_2)^2} < \frac{q_1 - \hat{q}_1(0)}{q_k - \hat{q}_k(0)}. \tag{C.20}$$

However, $f(x) = (q_1 - x)/(q_k - x)$ is increasing with $x \in [0, q_k]$. Therefore, (C.19) and (C.20) lead to a contradiction from the fact that $\bar{q}_1 < \bar{q}_2$.

In case (ii), multiple crossing implies that there exists at least two switching points, $s_1 < s_2$, such that $\hat{q}_k(t) > \hat{q}_1(t)$ if and only if $t \in (s_1, s_2)$. First, suppose that $q_1 - \hat{q}_1(0) < q_k - \hat{q}_k(0)$. Recall that $\dot{\hat{q}}_j(t) = \frac{\dot{B}_j(t)(q_j - \hat{q}_j(t))^2}{B_0(q_j - \hat{q}_j(0))}$ for any $j = 1, \dots, K$. Under the greedy policy, $\dot{B}_1(t) > \dot{B}_k(t)$ for $t < s_1$ because $\dot{B}_j(t) = d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))$ and $\hat{q}_1(t) > \hat{q}_k(t)$ and $g(z_1(t)) < g(z_k(t))$. Combined with the fact that $q_1 - \hat{q}_1(s_1) = q_1 - \bar{q}_1 > q_k - \bar{q}_1 = q_k - \hat{q}_k(s_1)$, we have $\dot{\hat{q}}_1(t) \geq \dot{\hat{q}}_k(t)$ for t sufficiently close to, but smaller than s_1. This leads to a contradiction because near the crossing point s_1, $\hat{q}_k(t)$ must be increasing faster than $\hat{q}_1(t)$.

Now, suppose that $q_1 - \hat{q}_1(0) \geq q_k - \hat{q}_k(0)$. Observe that

$$\begin{aligned} \dot{\hat{q}}_1(s_1) \leq \dot{\hat{q}}_k(s_1) &\Longleftrightarrow \frac{\dot{B}_1(s_1)(q_1 - \bar{q}_1)^2}{q_1 - \hat{q}_1(0)} \leq \frac{\dot{B}_k(s_1)(q_k - \bar{q}_1)^2}{q_k - \hat{q}_k(0)} \\ &\Longrightarrow \frac{(q_1 - \bar{q}_1)^2}{(q_k - \bar{q}_1)^2} \leq \frac{q_1 - \hat{q}_1(0)}{q_k - \bar{q}_k(0)}, \end{aligned} \tag{C.21}$$

where the second line follows from the fact that $\dot{B}_1(s_1) \geq \dot{B}_k(s_1)$. However, observe also that

$$\frac{(q_1 - \bar{q}_1)^2}{(q_k - \bar{q}_1)^2} > \frac{q_1 - \bar{q}_1}{q_k - \bar{q}_1} > \frac{q_1 - \hat{q}_1(0)}{q_k - \hat{q}_1(0)} > \frac{q_1 - \hat{q}_1(0)}{q_k - \hat{q}_k(0)}, \tag{C.22}$$

where the first inequality follows from the fact that $(q_1 - \bar{q}_1)/(q_k - \bar{q}_1) > 1$, the second follows from the fact that $f(x) = (q_1 - x)/(q_k - x)$ is increasing with $x \in [0, q_k]$, and the third follows from the construction that $\hat{q}_1(0) > \hat{q}_k(0)$. Comparing (C.21) and (C.2.2) leads to a contradiction. This completes the proof of the lemma. □

For $j \geq 2$, define $s_j := \inf\{t \geq 0 : \hat{q}_1(t) \geq \hat{q}_j(t)\}$. If $\hat{q}_{1,0} < \hat{q}_{j,0}$, then there exists a unique crossing between $\hat{q}_1(t)$ and $\hat{q}_j(t)$ by Lemma C.3 and $s_j \in (0, \infty)$. If $\hat{q}_{1,0} \geq \hat{q}_{j,0}$, then we have $s_j = 0$.

LEMMA C.4. *Suppose that Assumption 4.1 holds and fix $j \geq 2$ such that $\hat{q}_{1,0} < \hat{q}_{j,0}$. Then, under the greedy policy, for each $j \geq 2$, $\hat{q}_1(s_j) \to q_j$ as $K \to \infty$.*

Proof of Lemma C.4. In this proof, some variables depend on the quality configuration Q^K but we suppress the dependence in function arguments for clarity of exposition. Without loss of generality we assume that $s_2 \geq s_3 \geq \cdots \geq s_K$; the proof can be easily extended to general cases by re-indexing products. In what follows, we consider two cases: (i) $\hat{q}_j(0) > q_j$ and (ii) $\hat{q}_j(0) < q_j$. (If $\hat{q}_j(0) = q_j$, then the proof would be trivial since $\hat{q}_j(t) = q_j$ for all $t \geq 0$ such that $\hat{q}_1(s_j) = q_j$.)

Case (i). Suppose that $\hat{q}_j(0) > q_j$. Towards a contradiction, assume that the $\liminf_{K\to\infty} \hat{q}_j(s_j) \geq q_j + \varepsilon$ for some $\varepsilon > 0$. From (3.3), it follows that as $K \to \infty$,

$$\limsup_{K\to\infty} B_j(s_j) = B_0 \frac{\hat{q}_j(0) - q_j}{\varepsilon}. \tag{C.23}$$

Since $\hat{q}_j(t) \in [0,1]$ for all j, observe further that

$$\begin{aligned}
B_j(s_j) &= \int_0^{s_j} \frac{e^{\hat{q}_j(t) - p - g(z_j(t))}}{1 + \sum_{l=1}^K e^{\hat{q}_l(t) - p - g(z_l(t))}} \, dt \\
&= s_j - \int_0^{s_j} \frac{1}{1 + \sum_{l=1}^K e^{\hat{q}_l(t) - p - g(z_l(t))}} \begin{pmatrix} 1 + \sum_{i<j} e^{\hat{q}_i(t) - p - g(z_i(t))} + \\ \sum_{i>j} e^{\hat{q}_i(t) - p - g(z_i(t))} \end{pmatrix} dt \\
&\leq s_j - \int_0^{s_j} \frac{1}{1 + \sum_{l=1}^K e^{\hat{q}_l(t) - p - g(z_l(t))}} \begin{pmatrix} 1 + \sum_{i<j} e^{-p - g(z_i(t))} + \\ \sum_{i>j} e^{-p - g(z_i(t))} \end{pmatrix} dt \\
&= s_j - \sum_{k=j}^K \int_{s_{k+1}}^{s_k} \frac{1}{1 + \sum_{l=1}^K e^{\hat{q}_l(t) - p - g(z_l(t))}} \begin{pmatrix} 1 + \sum_{i<k} e^{-p - g(i-1)} + \\ \sum_{i>k} e^{-p - g(i)} \end{pmatrix} dt \\
&= s_j - \frac{B_0(q_1 - \hat{q}_{1,0})}{v_1} \sum_{k=j}^K \psi(\hat{q}_1(s_{k+1}) - q_1, \hat{q}_1(s_k) - q_1) \begin{pmatrix} 1 + \sum_{i<k} e^{g(k) - g(i-1)} + \\ \sum_{i>k} e^{g(k) - g(i)} \end{pmatrix},
\end{aligned} \tag{C.24}$$

where the inequality follows by replacing $\hat{q}_i(t) \in [0,1]$ with zero in the numerator of the integrand. We deduce that

$$\frac{s_j}{B_0} \geq B_j(s_j) + \frac{q_1 - \hat{q}_{1,0}}{v_1} \sum_{k=j}^K \psi(\hat{q}_1(s_{k+1}) - q_1, \hat{q}_1(s_k) - q_1) \begin{pmatrix} 1 + \sum_{i<k} e^{g(k) - g(i-1)} + \\ \sum_{i>k} e^{g(k) - g(i)} \end{pmatrix}.$$

Note that the last term in the preceding equation can be bounded as

$$\begin{pmatrix} 1 + \sum_{i<k} e^{g(k) - g(i-1)} + \\ \sum_{i>k} e^{g(k) - g(i)} \end{pmatrix} \geq \sum_{i=2}^{k-1} e^{g(k) - g(i)} \geq e^{g(k)} - 1, \tag{C.25}$$

where the first inequality follows from the fact that $g(i) > g(i-1)$ and by dropping positive terms. To find a lower bound of s_j in a more tractable form, observe that

$$\begin{aligned}
&\sum_{k=j}^{K} \psi(\hat{q}_1(s_{k+1}) - q_1, \hat{q}_1(s_k) - q_1) \begin{pmatrix} 1 + \sum_{i<k} \mathrm{e}^{g(k)-g(i-1)} + \\ \sum_{i>k} \mathrm{e}^{g(k)-g(i)} \end{pmatrix} \\
\geq\ &\sum_{k=j}^{K} \psi(\hat{q}_1(s_{k+1}) - q_1, \hat{q}_1(s_k) - q_1)(\mathrm{e}^{g(k)} - 1) \qquad \text{(C.26)} \\
\geq\ &\frac{\mathrm{e}^{-\hat{q}_{1,0}-q_1}}{(\hat{q}_{1,0}-q_1)^2} \sum_{k=j}^{K} (\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}))(\mathrm{e}^{g(k)} - 1),
\end{aligned}$$

where the last inequality follows from $\psi(x_1, x_2) = \int_{x_1}^{x_2} \mathrm{e}^{-y}/y^2 dy \geq \mathrm{e}^{-x_2} \int_{x_1}^{x_2} 1/y^2 dy$ for $x_1 < x_2 < 0$. Furthermore, there exists a constant c such that $\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) \geq c/K$ for all $k \leq K$; otherwise, $\sum_{k=j}^{K} (\hat{q}_1(s_k) - \hat{q}_1(s_{k+1})) = \hat{q}_1(s_j) - \hat{q}_1(0) \to 0$ as $K \to \infty$, but this would lead to a contradiction to the fact that $\hat{q}_1(s_j) - \hat{q}_1(0) > q_j - \hat{q}_1(0) > 0$ by Assumption 4.1(c). Let $G_1(n) \coloneqq \sum_{k=1}^{n} \mathrm{e}^{-g(k)}$ and $G_2(n) \coloneqq \sum_{k=1}^{n} \mathrm{e}^{g(k)}$. The preceding observations imply that $s_j = \Omega(G_2(K)/K)$ as $K \to \infty$. Furthermore, since $z_j(t) \leq j$ for $t \leq s_j$ and $\hat{q}_j(t) \in [0,1]$, it can be seen that there exists a constant $c' > 0$ that is independent of K such that

$$\tilde{d}_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) = \frac{\mathrm{e}^{\hat{q}_j(t)-p-g(z_j(t))}}{1 + \sum_{l=1}^{K} \mathrm{e}^{\hat{q}_l(t)-p-g(z_l(t))}} \geq \frac{c'}{1 + \mathrm{e}^{1-p} G_1(K)}, \qquad \text{(C.27)}$$

from which we deduce that $\tilde{d}_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) = \Omega(1/G_1(K))$ as $K \to \infty$ for any $\boldsymbol{z}$, and therefore, we have

$$B_j(s_j) = \int_0^{s_j} \tilde{d}_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\, \mathrm{d}t \geq \Omega\left(\frac{G_2(K)}{K G_1(K)}\right) \to \infty \quad \text{as} \quad K \to \infty, \qquad \text{(C.28)}$$

where the inequality follows from (C.27) and the fact that $s_j = \Omega(G_2(K)/K)$ as $K \to \infty$ and the limit follows from the definitions of $G_1(\cdot)$ and $G_2(\cdot)$. This leads to a contradiction to the assumption that $\limsup_{K\to\infty} B_j(s_j) \leq B_0(\hat{q}_j(0) - q_j)/\varepsilon < \infty$ for some $\varepsilon > 0$.

Case (ii). Now suppose that $\hat{q}_j(0) < q_j$. Note that if $\hat{q}_j(0) \leq \hat{q}_1(0)$, then $\hat{q}_j(t) \leq \hat{q}_1(t)$ for all t such that there is no crossing point between $\hat{q}_1(t)$ and $\hat{q}_j(t)$. Hence, it suffices to consider $\hat{q}_j(0) \in (\hat{q}_1(0), q_j)$. Note that, without loss of generality, we may assume that $\hat{q}_j(0) \geq \hat{q}_1(0) + \delta$ for an arbitrarily small $\delta > 0$ for each configuration Q^K, $K \geq 2$. (If this is violated for some K, then we may consider a subsequence $K_1 < K_2 < \cdots$ for which the condition $\hat{q}_j(0) \geq \hat{q}_1(0) + \delta$ holds for all K_m, $m \in \mathbb{N}$.) The rest of the proof follows from the same logical steps as *Case (i)*. To avoid repetitions, we only remark that in (C.26), there exists a constant $c'' > 0$ such that $\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) \geq c''/K$ for all $k \leq K$; otherwise, $\sum_{k=j}^{K} \hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) = \hat{q}_1(s_j) - \hat{q}_1(0) \to 0$ as $K \to \infty$, which violates the assumption that $\hat{q}_1(s_j) \geq \hat{q}_j(0) \geq \hat{q}_1(0) + \delta$. This implies that $s_j = \Omega(G_2(K)/K)$ as $K \to \infty$. Using (C.27) and (C.28) once again, we derive a contradiction to the assumption that $B_j(s_j) < \infty$. This concludes the proof of the lemma. □

Proof of Proposition 4.2. For simple exposition, we fix $Q^K \in Q^K(\eta)$ and suppress the argument η. For the greedy policy, the revenue over time horizon $[0,T]$ can be characterized as

$$\begin{aligned}\frac{R_T^{\mathsf{G}}}{p} &= \int_0^T \sum_{j=1}^K d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t = T - \int_0^T 1 - \sum_{j=1}^K d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t \\ &= T - \begin{pmatrix} \sum_{j=2}^K \int_{s_{j+1}}^{s_j} \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(j)}\,\mathrm{d}t \\ +\int_{s_2}^T \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(1)}\,\mathrm{d}t \end{pmatrix} \\ &= T - \frac{B_0}{v_1}(q_1 - \hat{q}_{1,0}) \begin{pmatrix} \sum_{j=2}^K \mathrm{e}^{g(j)}\psi(\hat{q}_1(s_{j+1}) - q_1, \hat{q}_1(s_j) - q_1) \\ +\mathrm{e}^{g(1)}\psi(\hat{q}_1(s_2) - q_1, \hat{q}_1(T) - q_1) \end{pmatrix}. \end{aligned} \tag{C.29}$$

For the optimal policy, using the similar logical steps, one can show that

$$\frac{R_T^*}{p} = T - \frac{B_0}{v_1}(q_1 - \hat{q}_{1,0}) \left(\mathrm{e}^{g(1)}\psi(\hat{q}_{1,0} - q_1, \hat{q}_1^*(T) - q_1)\right). \tag{C.30}$$

Combining these expressions, we deduce

$$\frac{R_T^* - R_T^{\mathsf{G}}}{p} = \frac{B_0(q_1 - \hat{q}_{1,0})}{v_1} \begin{pmatrix} \sum_{j=2}^K \mathrm{e}^{g(j)}\psi(\hat{q}_1(s_{j+1}) - q_1, \hat{q}_1(s_j) - q_1) \\ +\mathrm{e}^{g(1)}\psi(\hat{q}_1(s_2) - q_1, \hat{q}_1(T) - q_1) \\ -\mathrm{e}^{g(1)}\psi(\hat{q}_{1,0} - q_1, \hat{q}_1^*(T) - q_1) \end{pmatrix}. \tag{C.31}$$

Recalling the definitions of the regret $\mathfrak{R}^{\mathsf{G}} := \lim_{T\to\infty}\{R_T^* - R_T^{\mathsf{G}}\}$, and using Lemma C.4, it follows that as $K \to \infty$,

$$\mathfrak{R}_{\mathsf{G}}(Q) \sim \frac{B_0 p(q_1 - \hat{q}_{1,0})}{v_1} \sum_{j=2}^K \left(\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}\right)\phi_j, \tag{C.32}$$

where ϕ_j is defined in (B.6).

Since the (asymptotic) regret, characterized in (C.32), does not depend on the prior belief $\hat{q}_{j,0}$, $j \neq 1$, it suffices to consider the quality configuration such that $\hat{q}_{j,0} = q_j$ for each $j \neq 1$. That is, we may restrict our attention to the set of quality configurations $\mathcal{Q}_1^K \subset \mathcal{Q}^K(\eta)$ such that

$$\mathcal{Q}_1^K = \left\{ (\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \middle| \begin{array}{c} q_1 < 1 \ \text{ and } \ q_1 - \eta = q_2 \geq q_3 \geq \cdots \geq q_K \geq 0 \\ \hat{q}_{1,0} = 0 \ \text{ and } \ \hat{q}_{j,0} = q_j \ \text{ for } \ j \geq 1 \end{array} \right\}. \tag{C.33}$$

To show that $\mathfrak{R}_{\mathsf{G}}(Q)$ for $Q \in \mathcal{Q}_1^K$ is increasing with $q_j \leq 1 - \eta$ for $j \neq 1$, consider $Q = (\boldsymbol{q}, \hat{\boldsymbol{q}}_0) \in \mathcal{Q}_1^K$ and $Q' = (\boldsymbol{q}', \hat{\boldsymbol{q}}_0') \in \mathcal{Q}_1^K$ such that $q_i \geq q_i'$ for all $i = 1, \ldots, K$ and $q_j > q_j'$ for some $j \neq 1$. From (C.32), under the configuration Q', the regret can be written as

$$\mathfrak{R}_{\mathsf{G}}(Q') \sim \frac{B_0(q_1 - \hat{q}_{1,0})}{v_1} \sum_{j=2}^K \left(\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}\right)\phi_j' \ \text{ as } \ K \to \infty, \tag{C.34}$$

where ϕ_j' is defined in (B.6) with q_j being replaced by q_j' for each $j \neq 1$. Furthermore, observe that

$$\sum_{j=2}^K \left(\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}\right)\phi_j - \sum_{j=2}^K \left(\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}\right)\phi_j' = \sum_{j=2}^K \left(\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}\right)\left(\int_{q_{j+1}}^{q_j} \frac{\mathrm{e}^{-y}}{y^2}\,\mathrm{d}y - \int_{q_{j+1}'}^{q_j'} \frac{\mathrm{e}^{-y}}{y^2}\,\mathrm{d}y\right) > 0,$$

where the inequality follows from the fact that $q_j \geq q'_j$ for each $j \neq 1$ with at least one strict inequality. Thus, we deduce that $\mathfrak{R}^{\mathsf{G}}(Q) > \mathfrak{R}^{\mathsf{G}}(Q')$. Finally, note that the limit in (C.32) is increasing with q_1. Therefore, we establish that for sufficiently large K, the regret in (C.32) is maximized in configuration $Q^K_*(\eta) \in \mathcal{Q}^K_1$. This concludes the proof of the proposition. □

Proof of Theorem 4.1. Define $\bar{\phi} := \psi(-q_1, 1-\eta-q_1)$, where the function ψ is defined in (4.12). Let s be the switching time such that $\hat{q}_1(t) \geq 1-\eta$ if and only if $t \geq s$. We fix $Q^K = \bar{Q}^K_*(\eta)$ defined in the statement of the theorem and suppress Q^K in function arguments for simple exposition. We prove the result in two steps. We first derive the expression for the regret in (4.13). Then, we prove the properties of the function $\mathfrak{M}^K_{\mathsf{G}}(\eta)$.

Step 1. To derive the revenue for the greedy policy, observe that

$$\frac{R^{\mathsf{G}}_T}{p} = \sum_{j=1}^{K} \int_0^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t = \underbrace{\int_0^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t}_{A_1} + \sum_{j=2}^{K} \underbrace{\int_0^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t}_{A_j}. \tag{C.35}$$

It can be checked that

$$A_1 = \int_0^s d_1(\hat{q}(t), z(t))\,\mathrm{d}t + \int_s^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t = B_0\frac{1-\eta}{\eta} + \int_s^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t. \tag{C.36}$$

Also, for $j \neq 1$, we deduce from Lemma C.1 that

$$\begin{aligned} A_j &= \int_0^s d_j(\hat{q}(t), z(t))\,\mathrm{d}t + \int_s^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t \\ &= \mathrm{e}^{q_j - p - g(j-1)} \int_0^s \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(K)}\,\mathrm{d}t + \int_s^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t \\ &= \frac{B_0\bar{\phi}}{v_1} v_j \mathrm{e}^{g(K)-g(j-1)} + \int_s^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t. \end{aligned} \tag{C.37}$$

Combining these into (C.35), we deduce

$$\frac{R^{\mathsf{G}}_T}{p} = B_0\frac{1-\eta}{\eta} + \frac{B_0\bar{\phi}}{v_1}\sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)} + \sum_{j=1}^{K}\int_s^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t. \tag{C.38}$$

To derive an expression for s, observe that

$$\begin{aligned} B_1(s) = B_0\frac{1-\eta}{\eta} &= \int_0^s \frac{\mathrm{e}^{\hat{q}_1(t)-p+g(K)}}{1+\sum_{n=1}^{K}\mathrm{e}^{\hat{q}_n(t)-p+z_n(t)}}\,\mathrm{d}t \\ &= s - \int_0^s \frac{1+\sum_{j=2}^{K}\mathrm{e}^{q_j-p-g(j-1)}}{1+\sum_{n=1}^{K}\mathrm{e}^{\hat{q}_n(t)-p-z_n(t)}}\,\mathrm{d}t \\ &= s - (1+\sum_{j=2}^{K}\mathrm{e}^{q_j-p-g(j-1)})\int_0^s \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(K)}\,\mathrm{d}t \\ &= s - (1+\sum_{j=2}^{K} v_j \mathrm{e}^{-g(j-1)})\frac{B_0\bar{\phi}}{v_1}\mathrm{e}^{g(K)}, \end{aligned} \tag{C.39}$$

from which we have

$$s = B_0 \frac{1-\eta}{\eta} + \left(1 + \sum_{j=2}^{K} v_j \mathrm{e}^{-g(j-1)}\right) \frac{B_0 \bar{\phi}}{v_1} \mathrm{e}^{g(K)}. \tag{C.40}$$

To derive the expression for the optimal revenue, one may follow the similar steps as above, from which it can be seen that

$$\frac{R_T^*}{p} = B_0 \frac{1-\eta}{\eta} + \frac{B_0 \bar{\phi}}{v_1} \sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)} + \sum_{j=1}^{K} \int_{s^*}^{T} d_j(\hat{q}^*(t), z^*(t))\, \mathrm{d}t. \tag{C.41}$$

Also, the switching time s^* can be written as

$$s^* = B_0 \frac{1-\eta}{\eta} + \left(1 + \sum_{j=2}^{K} v_j \mathrm{e}^{-g(j)}\right) \frac{B_0 \bar{\phi}}{v_1} \mathrm{e}^{g(1)}. \tag{C.42}$$

Finally, since $s^* \le s$ and $\tilde{d}_j(\hat{q}(t), \boldsymbol{z}_\infty) = \tilde{d}_j(\hat{q}^*(t-s+s^*), \boldsymbol{z}_\infty)$ for all $t \ge s$, it follows that the revenue under the greedy policy during $[s, s+u]$ is identical to the revenue under the optimal policy during $[s^*, s^*+u]$ for any $u > 0$. Furthermore, observe from (C.40) that $s \to \infty$ as $K \to \infty$. Therefore, it can be easily seen that $\hat{q}_j(t) \sim \hat{q}_j^*(t) \sim q_j$ for each $j \neq 1$ as $K \to \infty$, and hence that the instantaneous revenue $p \sum_{j=1}^{K} d_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t)) \sim p \sum_{j=1}^{K} d_j(\hat{\boldsymbol{q}}^*(t), \boldsymbol{z}^*(t)) \sim r_\infty$ as $K \to \infty$ for $t \ge s$. (The constant r_∞ depends on K, although we suppress the dependence in notation.) Recall the definition of the regret $\mathfrak{R}^{\mathsf{G}} := \lim_{T\to\infty}\{R_T^* - R_T^{\mathsf{G}}\}$ and it can be seen that, as $K \to \infty$,

$$\begin{aligned}
\frac{\mathfrak{R}_{\mathsf{G}}}{B_0/v_1} &\sim r_\infty(s - s^*) + p\bar{\phi} \sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)} - p\bar{\phi} \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)} \\
&= r_\infty \bar{\phi} \begin{pmatrix} \left(1 + \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right) - \\ \left(1 + \sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)}\right) \end{pmatrix} + p\bar{\phi} \begin{pmatrix} \sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)} - \\ \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)} \end{pmatrix} \\
&= \bar{\phi} \begin{pmatrix} r_\infty(\mathrm{e}^{g(K)} - \mathrm{e}^{g(1)}) + \\ (p - r_\infty)\left(\sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)} - \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)}\right) \end{pmatrix},
\end{aligned} \tag{C.43}$$

where the first equality follows from (C.40) and (C.42). Using the fact that $r_\infty = (p - r_\infty) \sum_{j=1}^{K} v_j \exp(-g(j))$, we further deduce that

$$\begin{aligned}
\frac{\mathfrak{R}^{\mathsf{G}}}{B_0/v_1} &\sim \bar{\phi}(p - r_\infty) \begin{pmatrix} (\mathrm{e}^{g(K)} - \mathrm{e}^{g(1)}) \sum_{j=1}^{K} v_j \mathrm{e}^{-g(j)} + \\ \sum_{j=2}^{K} v_j \mathrm{e}^{g(1)-g(j)} - \sum_{j=2}^{K} v_j \mathrm{e}^{g(K)-g(j-1)} \end{pmatrix} \\
&= \bar{\phi}(p - r_\infty) \begin{pmatrix} \sum_{j=2}^{K} v_j (\mathrm{e}^{g(K)-g(j)} - \mathrm{e}^{g(K)-g(j-1)}) \\ + v_1 \mathrm{e}^{g(K)-g(1)} - v_1 \end{pmatrix} \\
&= \bar{\phi}(p - r_\infty)\left(\mathrm{e}^{g(K)} - 1\right)(v_1 - v_2),
\end{aligned} \tag{C.44}$$

where the last equality follows from the fact that $g(1) = 0$ and $v_2 = v_3 = \cdots = v_K$. Therefore, we obtain the desired expression by letting $\mathfrak{M}_{\mathsf{G}}^K(\eta) := \bar{\phi}(p - r_\infty)(v_1 - v_2)$.

Step 2. We show the properties of $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ stated in the theorem. For simplicity, let $h_1 := \mathrm{e}^{-p-g(K)}$ and $h_2 := \sum_{j\neq 1} \mathrm{e}^{-p-g(j-1)}$. To show that $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ is quasi-concave in η, observe that

$$r_\infty = p\frac{\sum_{j=1}^K \mathrm{e}^{q_j-p-g(j-1)}}{1+\sum_{j=1}^K \mathrm{e}^{q_j-p-g(j-1)}} = p\left(1-\frac{1}{1+\sum_{j=1}^K \mathrm{e}^{q_j-p-g(j-1)}}\right), \tag{C.45}$$

from which we deduce that $p-r_\infty = p/(1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2)$. From the definition of $\bar\phi$, it is easy to see that $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ is non-negative, continuous, and equal to zero for $\eta=0,1$. To show the quasi-concavity, recalling the expression $\mathfrak{M}_{\mathsf{G}}^K(\eta) = \bar\phi(p-r_\infty)(v_1-v_2)$, it suffices to check that

$$\tilde M^K(\eta) := \frac{\mathrm{e}-\mathrm{e}^{1-\eta}}{1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2}\bar\phi \tag{C.46}$$

is quasi-concave for $\eta\in[0,1]$. To this end, observe that

$$\frac{\partial \tilde M^K(\eta)}{\partial\eta} = -\frac{\mathrm{e}^{1-\eta}(1+\mathrm{e}h_1+\mathrm{e}h_2)}{(1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2)^2}\left(\frac{(\mathrm{e}-\mathrm{e}^{1-\eta})(1+\mathrm{e}h_1+\mathrm{e}h_2)}{(1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2)\mathrm{e}^{1-\eta}}\frac{\mathrm{e}^\eta}{(1-\eta)^2}-\bar\phi\right) \tag{C.47}$$

Note that the sign of $\partial\tilde M^K(\eta)/\partial\eta$ is determined by that of

$$Z^K(\eta) := \frac{(\mathrm{e}-\mathrm{e}^{1-\eta})(1+\mathrm{e}h_1+\mathrm{e}h_2)}{(1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2)\mathrm{e}^q}\frac{\mathrm{e}^\eta}{(1-\eta)^2}-\bar\phi. \tag{C.48}$$

After some straightforward algebra, one can deduce that

$$\frac{\partial Z^K(\eta)}{\partial\eta} = \frac{\mathrm{e}^\eta}{\eta^3}\left(\frac{2(1+\mathrm{e}h_1+\mathrm{e}^{1-\eta}h_2)(\mathrm{e}^{1-\eta}-q\mathrm{e})}{(1+\mathrm{e}h_1+\mathrm{e}h_2)\mathrm{e}^{1-\eta}}\right). \tag{C.49}$$

Since $\mathrm{e}^{1-\eta}-q\mathrm{e}>0$ for all $\eta\in[0,1]$, we deduce that $\partial Z^K(\eta)/\partial\eta>0$. Since $Z^K(0)<0$, we obtain that there exists some $\bar\eta\in(0,1)$ such that $Z^K(\eta)<0$ if and only if $\eta<\bar\eta$. This, in turn, implies that $\partial\tilde M^K(\eta)/\partial\eta>0$ if and only if $\eta<\bar\eta$. These observations imply that $\tilde M^K(\eta)$ is quasi-concave in $\eta\in[0,1]$, which completes the proof of the theorem. □

Proof of Corollary 4.1. The proof of the corollary follows from the expression of the regret in (4.13) and will be omitted. □

Proof of Corollary 4.2. Recall from the proof of Theorem 4.1 that maximizing $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ over $\eta\in[0,1]$ is equivalent to maximizing $\tilde M^K(\eta)$ defined in (C.46). Note that $\tilde M^K(\eta)\le(\mathrm{e}-\mathrm{e}^{1-\eta})\bar\phi$ and $\bar\phi=\psi(-1,-\eta)$ does not depend on K. Therefore, it is easy to check that $\tilde M^K(\eta)$ is uniformly bounded in the sense that there exists a constant $\bar M<\infty$ such that $\tilde M^K(\eta)\le\bar M$ for any $\eta\in[0,1]$ and $K\ge2$. This implies that the sequence $\{\tilde M^K(\cdot):K\ge2\}$ is uniformly convergent to $M^\infty(\cdot)$, and therefore, the maximizer $\eta_*^K=\arg\max_{\eta\in[0,1]}\{\tilde M^K(\eta)\}$ also converges to $\eta_*^\infty=\arg\max_{\eta\in[0,1]}\{\tilde M^\infty(\eta)\}$ as $K\to\infty$; that is,

$$\frac{\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta_\infty^*))}{\mathfrak{R}_{\mathsf{G}}(Q_*^K(\eta_K^*))}\to 1 \quad \text{as } K\to\infty. \tag{C.50}$$

Further, we have from Proposition 4.2 that as $K \to \infty$,

$$\begin{aligned}\frac{\mathfrak{R}_{\mathsf{G}}^{K+1}}{\mathfrak{R}_{\mathsf{G}}^{K}} &\sim \frac{\mathfrak{R}_{\mathsf{G}}(Q_*^{K+1}(\eta_*^{K+1}))}{\mathfrak{R}_{\mathsf{G}}(Q_*^{K}(\eta_*^{K}))} \\ &= \frac{\mathfrak{R}_{\mathsf{G}}(Q_*^{K+1}(\eta_*^{\infty}))}{\mathfrak{R}_{\mathsf{G}}(Q_*^{K}(\eta_*^{\infty}))} \frac{\mathfrak{R}_{\mathsf{G}}(Q_*^{K+1}(\eta_*^{K+1}))}{\mathfrak{R}_{\mathsf{G}}(Q_*^{K+1}(\eta_*^{\infty}))} \frac{\mathfrak{R}_{\mathsf{G}}(Q_*^{K}(\eta_*^{\infty}))}{\mathfrak{R}_{\mathsf{G}}(Q_*^{K}(\eta_*^{K}))}.\end{aligned} \tag{C.51}$$

The second and third terms on the right-hand side of the preceding equation converges to one as $K \to \infty$ by (C.50). Thus, the desired conclusion follows by applying Theorem 4.1, which concludes the proof of the theorem. □

C.2.3. Proofs for Section 4.4. As in Section C.2.2, we use the asymptotics with respect to the number of product K for proofs in this section. Hence, most variables depend on K, but we suppress the dependence to improve clarity. Moreover, we adopt the notation $s_{K+1} \le s_K \le \cdots \le s_2$ introduced in Section C.2.2. The following lemmas will be useful in proving Theorem 4.2.

LEMMA C.5. *Suppose that Assumption 4.1(a)-(b) hold. Fix a quality configuration $Q^K \in \mathcal{Q}^K$ and consider $k \le K$. Then, under the semi-greedy policy with $u \le \bar{u}$, there exists $s > 0$ such that (i) $\tilde{q}_1(s) = \tilde{q}_k(s)$, (ii) $\tilde{q}_1(t) < \tilde{q}_k(t)$ for all $t < s$, and (iii) $\tilde{q}_1(t) > \tilde{q}_k(t)$ for all $t > s$.*

Proof of Lemma C.5. Note that the implied belief process $\tilde{q}_k(t)$ satisfies $\dot{\tilde{q}}_k(t) = \frac{\dot{B}_k(t)}{B_k(t)+B_{k,0}}(q_k - \tilde{q}_k(t))$, just like the original belief process which satisfies $\dot{\hat{q}}_k(t) = \frac{\dot{B}_k(t)}{B_k(t)+B_{k,0}}(q_k - \hat{q}_k(t))$. For this reason, the proof of the lemma follows immediately from that of Lemma C.3 with the belief process $\{\hat{q}_k(t) : t \ge 0\}$ replaced by $\{\tilde{q}_k(t) : t \ge 0\}$. We omit the detail of the proof to avoid repetition. □

LEMMA C.6. *Suppose that Assumption 4.1 holds. Fix $j \ge 2$ such that $\tilde{q}_1(0) < \tilde{q}_j(0)$. Then, under the semi-greedy policy with $u \le \bar{u}$, $\tilde{q}_1(s_j) \to q_j$ as $K \to \infty$ for each $j \ge 2$.*

Proof of Lemma C.6. In this proof, we assume that $\hat{q}_j(0) > q_j$; the case with $\hat{q}_j(0) < q_j$ follows from identical steps and will be omitted to avoid repetition. Moreover, some variables depend on the quality configuration Q^K but we suppress the dependence for clarity of exposition. Define $s_j := \inf\{t \ge 0 : \tilde{q}_1(t) \ge \tilde{q}_j(t)\}$, which is well defined by Lemma C.5. Without loss of generality, we let $s_2 \ge s_3 \ge \cdots \ge s_K$; the proof can be extended to general cases by re-indexing products.

Towards a contradiction, assume that $\liminf_{K\to\infty} \tilde{q}_j(s_j) \ge q_j + \varepsilon$ for a sufficiently small $\varepsilon > 0$. From (3.3) and the fact that $\tilde{q}_j(t) = \hat{q}_j(t) + u/(B_j(t) + B_0)$, it follows that

$$\limsup_{K\to\infty} B_j(s_j) = \frac{B_0(\hat{q}_j(0) - q_j) + u}{\varepsilon}. \tag{C.52}$$

Taking $K \to \infty$ on both sides of inequality in (C.24), we deduce that

$$\begin{aligned}\frac{s_j}{B_0} \ge{}& \frac{(\hat{q}_j(0) - q_j) + u/B_0}{\varepsilon} \\ &+ \frac{q_1 - \hat{q}_{1,0}}{v_1} \sum_{k=j}^{\infty} \psi(\hat{q}_1(s_{k+1}) - q_1, \hat{q}_1(s_k) - q_1) \left(1 + \frac{\sum_{i<k} \mathrm{e}^{g(k)-g(i-1)}}{\sum_{i>k} \mathrm{e}^{g(k)-g(i)}} + \right).\end{aligned} \tag{C.53}$$

We can find a constant c such that $\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) \geq c/K$ for all $k \leq K$ and $K \geq 2$; otherwise, $\sum_{k=j}^{K}(\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) = \hat{q}_1(s_j) - \hat{q}_1(0) \to 0$ as $K \to \infty$. This leads to a contradiction because $\hat{q}_1(s_j) - \hat{q}_{1,0} > q_j - \hat{q}_{1,0}$ and q_j is the jth largest quality in configuration Q^K, which can only increase with K. Now, from (C.26) and the fact that $\hat{q}_1(s_k) - \hat{q}_1(s_{k+1}) \geq c/K$ for all $k \leq K$ and $K \geq 2$, the third term on the right-hand side of (C.53) increases to ∞ as $K \to \infty$. This implies that $s_j = \Omega(G_2(K)/K)$ as $K \to \infty$, where $G_2(n) \coloneqq \sum_{k=1}^{n} \mathrm{e}^{g(k)}$. Furthermore, $\tilde{d}_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}) = \Omega(1/G_1(K))$ by (C.27), where $G_1(n) \coloneqq \sum_{k=1}^{n} \mathrm{e}^{-g(k)}$. Therefore, we have

$$B_j(s_j) = \int_0^{s_j} \tilde{d}_j(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t))\,\mathrm{d}t \geq \Omega\left(\frac{G_2(K)}{KG_1(K)}\right) \to \infty \quad \text{as} \quad K \to \infty, \tag{C.54}$$

where the limit follows from the definitions of $G_1(\cdot)$ and $G_2(\cdot)$. This leads to a contradiction to the assumption that $B_j(s_j) = (B_0(\hat{q}_j(0) - q_j) + u)/\varepsilon < \infty$ for fixed $\varepsilon > 0$. This concludes the proof of the lemma. □

Proof of Theorem 4.2. We prove the theorem in three steps. First, we prove that $Q_*^K(\eta)$ is asymptotically the worst-case configuration in $\mathcal{Q}^K(\eta)$ as $K \to \infty$. Second, we characterize the regret in the configuration $Q_*^K(\eta)$ to (4.20). Third, we prove the properties of the function $\mathfrak{M}_{\mathsf{SG}}^K(\eta, u)$ defined in (4.20).

Step 1. We show that $Q_*^K(\eta)$ is a worst-case configuration in $\mathcal{Q}^K(\eta)$ asymptotically as $K \to \infty$. The proof of this step follows from exactly the same logical steps of that of Proposition 4.2. Concretely, note that the expression of the revenue under the semi-greedy policy is equivalent to the right-hand side of (C.29). Thus, from (C.32), it can be seen that as $K \to \infty$,

$$\mathfrak{R}_{\mathsf{SG}}(Q) \sim \frac{pB_0}{v_1}(q_1 - \hat{q}_{1,0}) \sum_{j=2}^{K} (\mathrm{e}^{g(j)} - \mathrm{e}^{g(1)}) \psi(\hat{q}_{1,0} - q_1, q_j - q_1), \tag{C.55}$$

where we use the fact that $\hat{q}_1(s_j) \to q_j$ as $K \to \infty$ (Lemma C.6). Since the regret characterized on the right-hand side of (C.55) does not depend on the prior belief $\hat{q}_{j,0}$, $j \neq 1$, it suffices to consider the quality configuration such that $\hat{q}_{j,0} = q_j$ for each $j \neq 1$. That is, we may restrict our attention to $\mathcal{Q}_1^K \subset \mathcal{Q}^K$, where $\mathcal{Q}_1^K$ is (C.33). Note that $\mathfrak{R}_{\mathsf{SG}}(Q)$ in (C.55) has the same functional form as $\mathfrak{R}_{\mathsf{G}}(Q)$ in (C.32). Thus, using the same logical steps as in the proof of Proposition 4.2, one can establish that $Q_*^K(\eta) \in \mathcal{Q}_1^K$ maximizes the regret for sufficiently large K.

Step 2. Fix $Q^K = Q_*^K(\eta)$. First, we derive the expression for the T-period revenue under the semi-greedy policy. Define $\tilde{\tau} \coloneqq \inf_t\{\tilde{q}_1(t) \geq q_2\}$ and $\tau \coloneqq \inf_t\{\hat{q}_1(t) \geq q_2\}$. By definition, $\tilde{\tau} \leq \tau$. Observe that

$$\frac{R_T^{\mathsf{SG}}}{p} = \sum_{j=1}^{K} \int_0^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t = \underbrace{\int_0^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t}_{A_1} + \sum_{j=2}^{K} \underbrace{\int_0^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t}_{A_j}. \tag{C.56}$$

For $k=1$, we have

$$A_1 = \int_0^\tau d_1(\hat{q}(t), z(t))\,\mathrm{d}t + \int_\tau^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t = B_0\frac{1-\eta}{\eta} + \int_\tau^T d_1(\hat{q}(t), z(t))\,\mathrm{d}t. \tag{C.57}$$

For $k \neq 1$, observe

$$A_j = \int_0^{\tilde{\tau}} d_j(\hat{q}(t), z(t))\,\mathrm{d}t + \int_{\tilde{\tau}}^\tau d_j(\hat{q}(t), z(t))\,\mathrm{d}t + \int_\tau^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t,$$

where

$$\begin{aligned}
\int_0^{\tilde{\tau}} d_j(\hat{q}(t), z(t))\,\mathrm{d}t &= \mathrm{e}^{q_j - p - g(j-1)} \int_0^{\tilde{\tau}} \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(K)}\,\mathrm{d}t \\
&= \frac{B_0 v_j}{v_1}\mathrm{e}^{g(K)-g(j-1)}\psi(\hat{q}_{1,0}-q_1, \hat{q}_1(\tilde{\tau}_1)-q_1) \\
\int_{\tilde{\tau}}^\tau d_j(\hat{q}(t), z(t))\,\mathrm{d}t &= \mathrm{e}^{q_j - p - g(j)} \int_{\tilde{\tau}}^\tau \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(1)}\,\mathrm{d}t \\
&= \frac{B_0 v_j}{v_1}\mathrm{e}^{g(1)-g(j)}\psi(\hat{q}_1(\tilde{\tau})-q_1, 1-\eta-q_1).
\end{aligned}$$

Furthermore, observe that at $t=\tilde{\tau}$,

$$q_1 + \frac{(\hat{q}_{1,0}-q_1)B_0+u}{B_1(\tilde{\tau})+B_0} = \hat{q}_1(\tilde{\tau}) \iff B_1(\tilde{\tau}) + B_0 = \frac{(\hat{q}_{1,0}-q_1)B_0+u}{\hat{q}_1(\tilde{\tau})-q_1}. \tag{C.58}$$

Lemma C.6 implies that $\tilde{q}_1(\tilde{\tau}) \to q_2 = 1-\eta$ as $K\to\infty$, from which we obtain that as $K\to\infty$,

$$\begin{aligned}
\hat{q}_1(\tilde{\tau}) &\to q_2 - \frac{u}{B_1(\tilde{\tau})+B_0} \\
&= q_2 - \frac{(1-\eta-q_1)u}{(\hat{q}_{1,0}-q_1)B_0+u}.
\end{aligned} \tag{C.59}$$

Combining these into (C.56), and recalling that $\hat{q}_{1,0}=0$ and $\hat{q}_1=1$ in configuration $Q_*^K(\eta)$, we deduce, as $K\to\infty$,

$$\begin{aligned}
\frac{R_T^{\mathsf{SG}}}{p} \sim B_0\frac{1-\eta}{\eta} &+ \sum_{j=1}^K \int_\tau^T d_j(\hat{q}(t), z(t))\,\mathrm{d}t \\
&+ \frac{B_0}{v_1}\sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j-1)}\psi\left(-1, -\frac{\eta B_0}{B_0-u}\right) \\
&+ \frac{B_0}{v_1}\sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)}\psi\left(-\frac{\eta B_0}{B_0-u}, -\eta\right).
\end{aligned} \tag{C.60}$$

To derive an expression for τ, observe that

$$\begin{aligned}
B_1(\tau) &= \int_0^\tau \frac{\mathrm{e}^{\hat{q}_1(t)-p+g(K)}}{1+\sum_{n=1}^K \mathrm{e}^{\hat{q}_n(t)-p+z_n(t)}}\,\mathrm{d}t \\
&= \tau - \int_0^{\tilde{\tau}} \frac{1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j-1)}}{1+\sum_{n=1}^K \mathrm{e}^{\hat{q}_n(t)-p-z_n(t)}}\,\mathrm{d}t - \int_{\tilde{\tau}}^\tau \frac{1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j)}}{1+\sum_{n=1}^K \mathrm{e}^{\hat{q}_n(t)-p-z_n(t)}}\,\mathrm{d}t.
\end{aligned}$$

Thus, the second and third terms satisfy, as $K \to \infty$,

$$\begin{aligned}
\int_0^{\tilde{\tau}} \frac{1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j-1)}}{1+\sum_{n=1}^K \mathrm{e}^{\hat{q}_n(t)-p-z_n(t)}}\,\mathrm{d}t &= (1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j-1)})\int_0^{\tilde{\tau}} \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(K)}\,\mathrm{d}t \\
&= (1+\sum_{j=2}^K v_j\mathrm{e}^{-g(j-1)})\frac{B_0\mathrm{e}^{g(K)}}{v_1}\psi\left(-1,-\frac{\eta B_0}{B_0-u}\right), \\
\int_{\tilde{\tau}}^{\tau} \frac{1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j)}}{1+\sum_{n=1}^K \mathrm{e}^{\hat{q}_n(t)-p-z_n(t)}}\,\mathrm{d}t &= (1+\sum_{j=2}^K \mathrm{e}^{q_j-p-g(j)})\int_{\tilde{\tau}}^{\tau} \dot{B}_1(t)\mathrm{e}^{-\hat{q}_1(t)+p+g(1)}\,\mathrm{d}t \\
&= (1+\sum_{j=2}^K v_j\mathrm{e}^{-g(j)})\frac{B_0\mathrm{e}^{g(1)}}{v_1}\psi\left(-\frac{\eta B_0}{B_0-u},-\eta\right).
\end{aligned}$$

Combined with the fact that $B_1(\tau) = B_0\frac{1-\eta}{\eta}$, we have that as $K \to \infty$,

$$\begin{aligned}
\tau \sim B_0\frac{1-\eta}{\eta} &+ (1+\sum_{j=2}^K v_j\mathrm{e}^{-g(j-1)})\frac{B_0\mathrm{e}^{g(K)}}{v_1}\psi\left(-1,-\frac{\eta B_0}{B_0-u}\right) \\
&+ (1+\sum_{j=2}^K v_j\mathrm{e}^{-g(j)})\frac{B_0\mathrm{e}^{g(1)}}{v_1}\psi\left(-\frac{\eta B_0}{B_0-u},-\eta\right).
\end{aligned} \tag{C.61}$$

To derive the expression for the optimal revenue, one may follow the similar steps as above, from which it can be seen that

$$\frac{R_T^*}{p} = B_0\frac{1-\eta}{\eta} + \frac{B_0\psi(-1,-\eta)}{v_1}\sum_{j=2}^K v_j\mathrm{e}^{g(1)-g(j)} + \sum_{j=1}^K \int_{\tau^*}^T d_j(\hat{q}^*(t), z^*(t))\,\mathrm{d}t. \tag{C.62}$$

Also, the switching time τ^* can be written as

$$\tau^* = B_0\frac{1-\eta}{\eta} + \left(1+\sum_{j=2}^K v_j\mathrm{e}^{-g(j)}\right)\frac{B_0\psi(-1,-\eta)}{v_1}\mathrm{e}^{g(1)}. \tag{C.63}$$

Since $\tau^* \le \tau$ and $\tilde{d}_j(\hat{q}(t), z_\infty) = \tilde{d}_j(\hat{q}^*(t-\tau+\tau^*), z_\infty)$ for all $t \ge \tau$, it follows that the revenue under the greedy policy during $[\tau, \tau+s]$ is identical to the revenue under the optimal policy during $[\tau^*, \tau^*+s]$ for any $s > 0$. Furthermore, observe from (C.61) that $\tau \to \infty$ as $K \to \infty$. Therefore, it can be easily seen that $\hat{q}_j(t) \sim \hat{q}_j^*(t) \sim q_j$ for each $j \ne 1$ as $K \to \infty$, and hence the instantaneous revenue $\sum_{j=1}^K \tilde{d}(\hat{\boldsymbol{q}}(t), \boldsymbol{z}(t)) \sim \sum_{j=1}^K \tilde{d}(\hat{\boldsymbol{q}}^*(t), \boldsymbol{z}^*(t)) \sim r_\infty$ as $K \to \infty$ for $t \ge \tau$. (The constant r_∞ depends on K, although we suppress the dependence in notation.) Recalling the definition of the regret $\mathfrak{R}^{\mathsf{SG}} =$

$\lim_{T\to\infty}\{R_T^* - R_T^{\mathsf{SG}}\}$, we have as $K\to\infty$,

$$\begin{aligned}
\frac{\mathfrak{R}_{\mathsf{SG}}^K}{B_0/v_1} &\sim r_\infty \begin{pmatrix} \left(\mathrm{e}^{g(K)} + \sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j-1)}\right)\psi\left(-1, -\frac{\eta B_0}{B_0-u}\right) + \\ \left(\mathrm{e}^{g(1)} + \sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)}\right)\psi\left(-\frac{\eta B_0}{B_0-u}, -\eta\right) - \\ \left(\mathrm{e}^{g(1)} + \sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)}\right)\psi(-1,-\eta) \end{pmatrix} \\
&\quad + p \begin{pmatrix} \psi(-1,-\eta)\sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)} - \\ \psi\left(-1, -\frac{\eta B_0}{B_0-u}\right)\sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j-1)} - \\ \psi\left(-\frac{\eta B_0}{B_0-u}, -\eta\right)\sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)} \end{pmatrix} \\
&= (p - r_\infty) \begin{pmatrix} \psi(-1,-\eta)\sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)} - \\ \psi\left(-1, -\frac{\eta B_0}{B_0-u}\right)\sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j-1)} - \\ \psi\left(-\frac{\eta B_0}{B_0-u}, -\eta\right)\sum_{j=2}^K v_j \mathrm{e}^{g(1)-g(j)} \end{pmatrix} + r_\infty \begin{pmatrix} \mathrm{e}^{g(K)}\psi\left(\hat{-}1, -\frac{\eta B_0}{B_0-u}\right) + \\ \mathrm{e}^{g(1)}\psi\left(-\frac{\eta B_0}{B_0-u}, -\eta\right) - \\ \mathrm{e}^{g(1)}\psi(-1,-\eta) \end{pmatrix}.
\end{aligned} \tag{C.64}$$

To simplify the expression above, note that $\tilde{r}_\infty = (p - \tilde{r}_\infty)\sum_{j=1}^K v_j \mathrm{e}^{-g(j)}$, from which we deduce that, as $K\to\infty$,

$$\begin{aligned}
\frac{\mathfrak{R}_{\mathsf{SG}}^K}{B_0/v_1} &\sim (p - r_\infty)\psi\left(-1, -\eta - \frac{\eta u}{B_0 - u}\right)\left(\sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j-1)} - \sum_{j=2}^K v_j \mathrm{e}^{g(K)-g(j)}\right) \\
&= (p - r_\infty)\psi\left(-1, -\eta - \frac{\eta u}{B_0 - u}\right)\left(\mathrm{e}^{g(K)} - 1\right),
\end{aligned} \tag{C.65}$$

where we use the fact that $\tilde{r}_\infty - r_\infty \to 0$ as $K\to\infty$.

Step 3. We prove the properties of the function $\mathfrak{M}_{\mathsf{SG}}^K(\eta)$. The non-negativity and continuity of $\mathfrak{M}_{\mathsf{SG}}^K(\eta)$ is trivial from the definition in (4.20). Moreover, note that $\mathfrak{M}_{\mathsf{SG}}^K(\eta)$ in (4.20) differs from $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ in (4.13) only by the terms $\psi(-1, -\frac{\eta B_0}{B_0-u})$ and $\psi(-1,-\eta)$. From the chain rule, we have

$$\frac{\partial}{\partial\eta}\psi\left(-1, -\frac{\eta B_0}{B_0-u}\right) = \frac{B_0}{B_0-u}\frac{\partial}{\partial\eta}\psi(-1,-\eta), \tag{C.66}$$

from which we deduce that $\mathfrak{M}_{\mathsf{SG}}^K(\eta)$ increasing in η if and only if $\mathfrak{M}_{\mathsf{G}}^K(\eta)$ is increasing. Combined with the fact that $\mathfrak{R}_{\mathsf{G}}^K(\eta)$ is quasi-concave, the preceding observation implies that $\mathfrak{R}_{\mathsf{SG}}^K(\eta)$ is also quasi-concave. Lastly, the relation $\mathfrak{R}_{\mathsf{SG}}^K(\eta) < \mathfrak{R}_{\mathsf{G}}^K(\eta)$ follows from the fact that $\psi(a,b) = \int_a^b \frac{\mathrm{e}^{-y}}{y^2}\,\mathrm{d}y$ and that $\psi(-1, -\frac{\eta B_0}{B_0-u})$ is the integral of the function $\frac{\mathrm{e}^{-y}}{y^2}$ over the smaller interval than $\psi(-1,-\eta)$. This completes the proof of the theorem. □

Appendix D: Derivation of the Fluid Approximation

The derivation generalizes to the multi-product case the approach of Crapis et al. (2017). Consider a sequence of systems indexed by $m = 1, 2, \ldots$, where the m-th system describes a market where consumers arrive according to a Poisson process with parameter $\Lambda^m = m\Lambda$. The symbols $L_k^m(t), D_k^m(t)$ and $B_k^m(t)$ denote the numbers of likes, of dislikes, and of purchases for product k

before time t in system m, respectively. The corresponding *scaled state variables* for product k are denoted by $\bar{I}_k^m(t) = (\bar{L}_k^m(t), \bar{D}_k^m(t)) \coloneqq (L_k^m(t)/m,\, D_k^m(t)/m)$ and $\bar{\boldsymbol{I}}^m(t) \coloneqq (\bar{I}_1^m(t), \bar{I}_2^m(t), \dots, \bar{I}_K^m(t))$ is the total information available at time t in the m-th system. The following lemma establishes that, if the arrival rate of consumers grows unbounded, there exist two deterministic processes $L_k(t)$ and $D_k(t)$ that approximate with arbitrary precision the scaled state variables $\bar{L}_k^m(t)$ and $\bar{D}_k^m(t)$. In particular, defining $B_k(t) \coloneqq L_k(t) + D_k(t)$ and $\hat{q}_k(t) \coloneqq (L_k(t) + L_{k,0})/(B_k(t) + B_{k,0})$, this deterministic approximation allows us to describe the learning trajectories as the continuous time solution $\hat{\boldsymbol{q}}(t) \coloneqq (\hat{q}_1(t), \hat{q}_2(t), \dots, \hat{q}_K(t))$ of the ODE (3.1).

LEMMA D.1. *For every $t > 0$ and every $k = 1, \dots, K$ we have $\sup_{0 \le s \le t} |\bar{L}_k^m(s) - L_k(s)| \to 0$ and $\sup_{0 \le s \le t} |\bar{D}_k^m(s) - D_k(s)| \to 0$ for $n \to \infty$ almost surely. Moreover, for all $k = 1, \dots, K$, the processes $L_k(t)$ and $D_k(t)$ are deterministic and satisfy the differential relations*

$$\dot{L}_k(t) = \Lambda \, \mathsf{P}(r_k(t) = L \mid \boldsymbol{I}(t)) = \Lambda \, d_k(\hat{\boldsymbol{q}}(t)) \, q_k, \tag{D.1}$$

$$\dot{D}_k(t) = \Lambda \, \mathsf{P}(r_k(t) = D \mid \boldsymbol{I}(t)) = \Lambda \, d_k(\hat{\boldsymbol{q}}(t)) \, (1 - q_k), \tag{D.2}$$

where $\boldsymbol{I}(t) \coloneqq (I_1(t), I_2(t), \dots, I_K(t))$ and $I_k(t) \coloneqq (L_k(t), D_k(t))$.

Proof of Lemma D.1. The proof consists in verifying the conditions of Theorem 2.2 of Kurtz (1977/78). First, observe that $\bar{\boldsymbol{I}}^m(t) \in \{z/m \mid z \in \mathbb{Z}_+^{2K}\}$, where $\mathbb{Z}_+^d$ denotes the d-dimensional integer lattice. To validate the remaining hypothesis of the theorem, we first need to show that the scaled number of likes and dislikes $\bar{L}_k^m(t)$ and $\bar{D}_k^m(t)$ can be expressed as a suitable Poisson processes with time-dependent rate, and then we must prove that the following inequalities hold

$$\gamma_k^L(x) \le \Gamma_1^L(1 + |x|), \quad \gamma_k^D(x) \le \Gamma_1^D(1 + |x|), \tag{D.3}$$

$$|\gamma_k^L(x) - \gamma_k^L(y)| \le \Gamma_2^L |x - y|, \quad |\gamma_k^D(x) - \gamma_k^L(y)| \le \Gamma_2^D |x - y|, \tag{D.4}$$

for some positive constants Γ_1^L, Γ_2^L, Γ_1^D, and Γ_2^D, and for all $k \in \mathcal{P}_K$ for all $x, y \in \mathbb{R}^{2K}$.
We define, for $k \in \mathcal{P}_K$, the following functions:

$$\gamma_k^L(\boldsymbol{I}(t)) \coloneqq \mathsf{P}(r_k(t) = L \mid \boldsymbol{I}(t)), \quad \gamma_k^D(\boldsymbol{I}(t)) \coloneqq \mathsf{P}(r_k(t) = D \mid \boldsymbol{I}(t)),$$

Furthermore, let A^m be a Poisson process with parameter Λ^m and let $N_k^L(a), N_k^D(a)$ be independent Poisson processes with arrival rate a. Then we can write:

$$\bar{L}_k^m(t) = \frac{1}{m} \int_0^t \mathbb{1}\{r_k(s) = L \mid \bar{\boldsymbol{I}}^m(s)\} \, \mathrm{d}A^m(s) = \frac{1}{m} N_k^L\Big(\Lambda^m \int_0^t \mathsf{P}(r_k(s) = L \mid \bar{\boldsymbol{I}}^m(s)) \, \mathrm{d}s\Big), \tag{D.5}$$

where in the last equality we used a Poisson thinning argument to replace the counting process of consumers who liked product k with a Poisson process whose arrival rate is proportional to the probability of observing a like for product k. Similarly, one can show that $\bar{D}_k^m(t) = \frac{1}{m} N_k^D\Big(\Lambda^m \int_0^t \mathsf{P}(r_k(s) = D \mid \bar{\boldsymbol{I}}^m(s)) \, \mathrm{d}s\Big)$.

It remains to prove the inequalities (D.3) and (D.4). Since γ_k^L and γ_k^D are probabilities, the inequalities in (D.3) hold with $\Gamma_1^L = \Gamma_1^D = 1$ for all $k \in \mathcal{P}_K$. Moreover, from (D.1), we observe that γ_k^L depends on $\boldsymbol{I}(t)$ only through the quality estimates $\hat{q}_1(t), \hat{q}_2(t), \ldots, \hat{q}_K(t)$. We now show that the quality estimate $\hat{q}_k(t) = (L_k(t) + L_{k,0})/(B_k(t) + B_{k,0})$ is Lipschitz continuous in $\boldsymbol{I}(t)$. In fact, since $\hat{q}_k(t)$ does not depend on $L_k(t)$ and $D_k(t)$ if $k \neq i$ it is trivially Lipschitz continuous in $I_k(t) = (L_k(t), D_k(t))$ if $k \neq i$. Moreover, defining $l_k(t) = (L_k(t) + L_{k,0})/(B_k(t) + B_{k,0})$ and $l'_k(t) = (L'_k(t) + L_{k,0})/(B'_k(t) + B_{k,0})$, for all $I_k(t) = (L_k(t), D_k(t))$ and $I'_k(t) = (L'_k(t), D'_k(t))$ we have

$$\begin{aligned}
\left|\frac{L_k(t)+L_{k,0}}{B_k(t)+B_{k,0}} - \frac{L'_k(t)+L_{k,0}}{B'_k(t)+B_{k,0}}\right| &= \left|\frac{(L_k(t)+L_{k,0})(D'_k(t)+D_{k,0}) - (L'_k(t)+L_{k,0})(D_k(t)+D_{k,0})}{(B_k(t)+B_{k,0})(B'_k(t)+B_{k,0})}\right| \\
&= \left|\frac{(L_k(t)+L_{k,0})(D'_k(t)+D_{k,0}) - (L'_k(t)+L_{k,0})(D'_k(t)+D_{k,0})}{(B_k(t)+B_{k,0})(B'_k(t)+B_{k,0})}\right. \\
&\quad \left. + \frac{(L'_k(t)+L_{k,0})(D'_k(t)+D_{k,0}) - (L'_k(t)+L_{k,0})(D_k(t)+D_{k,0})}{(B_k(t)+B_{k,0})(B'_k(t)+B_{k,0})}\right| \\
&\leq \frac{1-l'_k(t)}{B_k(t)+B_{k,0}}|L_k(t) - L'_k(t)| + \frac{l'_k(t)}{B_k(t)+B_{k,0}}|D_k(t) - D'_k(t)| \\
&\leq \frac{2}{B_{k,0}}|I_k(t) - I'_k(t)|,
\end{aligned}$$

for all $k \in \mathcal{P}_K$. Noticing that $d_k(\cdot) \in C^\infty([0,1]^K)$ and that $[0,1]^K$ is trivially a compact convex set, we conclude that γ_k^L is Lipschitz continuous for all $k \in \mathcal{P}_K$. A similar proof can be provided for γ_k^D for all $k \in \mathcal{P}_K$, so this proves the inequalities in (D.4).

To conclude the proof, observe

$$\dot{\hat{q}}_k(t) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{L_k(t)+L_{k,0}}{B_k(t)+B_{k,0}} = \frac{\dot{L}_k(t)}{B_k(t)+B_{k,0}} - \frac{(L_k(t)+L_{k,0})\dot{B}_k(t)}{(B_k(t)+B_{k,0})^2} = \frac{\dot{B}_k(t)[q_k - \hat{q}_k(t)]}{B_k(t)+B_{k,0}},$$

from which we derive (3.1). □