



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Diffusion Approximations for a Class of Sequential Experimentation Problems

Victor F. Araman, René A. Caldentey

To cite this article:

Victor F. Araman, René A. Caldentey (2022) Diffusion Approximations for a Class of Sequential Experimentation Problems. Management Science 68(8):5958-5979. <https://doi.org/10.1287/mnsc.2021.4195>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Diffusion Approximations for a Class of Sequential Experimentation Problems

Victor F. Araman,^a René A. Caldentey^b

^aOlayan School of Business, American University of Beirut, Beirut 1107 2020, Lebanon; ^bBooth School of Business, The University of Chicago, Chicago, Illinois 60637

Contact: va03@aub.edu.lb,  <https://orcid.org/0000-0002-3583-8124> (VFA); rene.caldentey@chicagobooth.edu,

 <https://orcid.org/0000-0002-6767-9770> (RAC)

Received: October 31, 2019

Revised: February 5, 2021; June 22, 2021

Accepted: June 23, 2021

Published Online in Articles in Advance:
December 28, 2021

<https://doi.org/10.1287/mnsc.2021.4195>

Copyright: © 2021 INFORMS

Abstract. A decision maker (DM) must choose an action in order to maximize a reward function that depends on the DM's action as well as on an unknown parameter Θ . The DM can delay taking the action in order to experiment and gather additional information on Θ . We model the problem using a Bayesian sequential experimentation framework and use dynamic programming and diffusion-asymptotic analysis to solve it. For that, we consider environments in which the average number of experiments that is conducted per unit of time is large and the informativeness of each individual experiment is low. Under such regimes, we derive a diffusion approximation for the sequential experimentation problem, which provides a number of important insights about the nature of the problem and its solution. First, it reveals that the problems of (i) selecting the optimal sequence of experiments to use and (ii) deciding the optimal time when to stop experimenting decouple and can be solved independently. Second, it shows that an optimal experimentation policy is one that chooses the experiment that maximizes the instantaneous volatility of the belief process. Third, the diffusion approximation provides a more mathematically malleable formulation that we can solve in closed form and suggests efficient heuristics for the non-asymptotic regime. Our solution method also shows that the complexity of the problem grows only quadratically with the cardinality of the set of actions from which the decision maker can choose. We illustrate our methodology and results using a concrete application in the context of assortment selection and new product introduction. Specifically, we study the problem of a seller who wants to select an optimal assortment of products to launch into the marketplace and is uncertain about consumers' preferences. Motivated by emerging practices in e-commerce, we assume that the seller is able to use a *crowd voting* system to learn these preferences before a final assortment decision is made. In this context, we undertake an extensive numerical analysis to assess the value of learning and demonstrate the effectiveness and robustness of the heuristics derived from the diffusion approximation.

History: Accepted by Omar Besbes, revenue management and market analytics.

Funding: V. F. Araman acknowledges the financial support of the university research board at the American University of Beirut Funding (University Research Board) [Project 25857]. R. A. Caldentey thanks the University of Chicago Booth School of Business for financial support.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mnsc.2021.4195>.

Keywords: sequential testing • experimentation • Bayesian demand learning • experiment design • optimal stopping • dynamic programming • crowdvoting

1. Introduction

This paper is concerned with the problem faced by a decision maker (DM) who must choose an action a from a finite set of available actions \mathcal{A} in order to maximize a reward function $\mathcal{R}(a, \Theta)$ that depends on the action a taken as well as on an unknown parameter Θ . Instead of selecting an action immediately, the DM has the option of postponing this decision in order to *experiment* and gather additional information about the true value of Θ . In this context, the decision maker needs to select the most effective sequence of experiments to implement through time as well as the

time when to stop these experiments and select a final action $a \in \mathcal{A}$.

A wide range of applications can be modeled using this general framework. For example, the DM can be a factory manager who needs to decide if a batch of production meets specific quality standards. For that the DM can sample items sequentially to measure their individual condition and, accordingly, extrapolate the quality assessment on the entire batch (Qiu 2014). Alternatively, the DM can be a pharmaceutical company conducting a sequence of clinical trials to evaluate the efficacy of some new drug or vaccine (Armitage et al.

2002). In yet another example, the decision maker can be an educational institution designing computerized adaptive testing systems to assess the level of proficiency of a cohort of examinees in a particular subject area (Bartroff et al. 2008, Finkelman 2008).

One particular application, which has served as our initial motivation for this paper, relates to the problem of assortment selection in the context of new product introduction. Launching new products into the marketplace offers great opportunities for companies to generate new revenue streams and increase sales. However, such endeavors represent risky bets as consumers' preferences are typically unknown and unsuccessful products are a major liability generating possibly great capital expenditure, early markdowns, serious goodwill cost, and loss of market share.¹ To mitigate these risks, companies seek to test the market's reaction (e.g., value for the price) to new products before launching decisions are made.

In general, experimentation can be expensive and difficult to conduct effectively and probably worth doing only seldomly. However, in many situations, this reality is now changing as companies are beginning to recognize the potential to *crowdsource* such market-testing activities. Online experimentation has been indeed growing exponentially in the last decade or so. Companies such as Uber, Netflix, Amazon, Microsoft, and many more² are aggressively implementing market experimentation through dedicated platforms with the objective of continuously improving the online experiences of their customers and inferring customer preferences. In the context of new product introduction, some companies have created *crowd-voting* platforms (e.g., Threadless.com),³ on which customers can vote for their favorite products among a menu of available options. By doing so, companies generate continuous feedback from the "crowd" at almost no cost except often for the lack of accuracy and veracity of the data gathered. In view of these challenges, an effective execution of a crowd-voting system involves deciding what is the best assortment of products to display to each individual voter in order to maximize the speed of learning as well as when to stop the experimentation process and decide which new products should be commercialized (Kohavi and Thomke 2017). Section 6 is devoted to this particular crowd-voting example, which we use to illustrate the methodology and results that we develop first in Section 4 for the general case.

Motivated by the operating conditions of many online experimentation platforms, our formulation of the DM's problem and the corresponding analysis are based on two important and distinctive features: (i) the time epochs at which experimentation is possible are driven by an exogenous point process that the DM does not control, and (ii) the average number of

experiments that can be conducted per unit of time is large, but the amount of information generated by each individual experiment is low. In the context of the crowd-voting example, the first assumption accounts for a stochastic arrival of viewers/voters to the platform website. As for the second feature, it depicts the high velocity at which data can be collected online and also captures the fact that such data are inherently more noisy and less reliable than when experiments are more targeted and carefully designed (e.g., focus groups or surveying experts). Under these conditions, we are able to use asymptotic analysis to derive a diffusion approximation for both the sequential experimentation problem and the underlying optimal stopping problem that the decision maker must solve. As we see, the diffusion model provides a number of important insights about the nature of the problem and its solution. First, it reveals that the problems of (i) selecting the optimal sequence of experiments and (ii) deciding the optimal time to stop experimenting decouple and can be solved independently. Second, it shows that an optimal experimentation policy is one that chooses the experiment that maximizes the instantaneous volatility of the belief process, a proxy of the learning process. This *maximum volatility principle* reduces dramatically the complexity of the dynamic experimentation selection problem and its solution. Third, the diffusion approximation also provides a more mathematically malleable formulation of the optimal stopping problem that we can solve in closed form. Interestingly, the computational complexity of the latter grows only quadratically with the cardinality of the set of actions \mathcal{A} ; in fact, we show that solving a problem with $|\mathcal{A}|$ actions is equivalent to solving a collection of $|\mathcal{A}| - 1$ problems each with only two actions.

In addition, we obtain from our diffusion approximations, heuristics policies for the moderate, nonasymptotic regime. These heuristics turn out to be extremely effective and robust as shown in our numerical analysis. Finally, as a by-product of our analysis of the crowd-voting example in Section 6, we derive diffusion approximations for a setting in which experimentation and learning are driven by the choices that voters make under a multinomial choice model (MNL). Given the popularity of the MNL model to represent consumer preferences, we believe that our approach to obtaining diffusion approximations can possibly be applied to a number of other applications beyond those discussed in this paper.

2. Related Literature

Our paper is related to two streams of literature. Methodologically, we contribute to the literature on hypothesis testing and sequential design of experiments initiated by Wald (1947) in the early 1940s. In

terms of applications, we contribute to the operations literature on assortment planning and demand learning (e.g., Caro and Gallien 2007, Kök et al. 2009).

Sequential analysis is concerned with the problem of effectively detecting the validity of a hypothesis through sequential sampling or tests. The sequential probability ratio test (SPRT) developed by Wald (1945) (see also Wald and Wolfowitz 1948) establishes that, under certain conditions, an optimal policy is determined by the first exit time of an appropriately defined likelihood ratio process from a bounded interval; the end points of this interval are determined by prespecified type I and II error targets. The initial formulation and ideas of Wald's (1945) SPRT test have been applied to a wide range of applications and extended in many different directions (e.g., Siegmund 1985, Lai 2001). One important extension relevant to our work relates to the problem of sequential design of experiments, in which the DM chooses dynamically the experiments to undertake from a set of available options (e.g., Robbins 1952; Chernoff 1959, 1972) and does that until the DM decides to stop and selects what the DM believes is the true hypothesis.

In terms of solution techniques, large sample analysis has been commonly used to study sequential hypothesis testing problems and evaluate the asymptotic optimality of concrete (often simple) policies. The asymptotic regime in many of these studies is obtained by assuming that the cost of experimentation goes to zero (e.g., Chernoff 1959, Keener 1984). Chernoff (1959, p. 757) mentions that "it may pay to continue sampling even though we are almost convinced about which is the true state of nature." The alluded "inefficiency" in Chernoff's (1959) regime is required to guarantee a probability of error that is proportional to the cost of experimentation that is becoming increasingly small. Our work also relies on a type of asymptotic analysis in which the number of experiments grows large; however, our approach differs significantly from large sample methods as we not only scale the number of experiments but simultaneously decrease the informativeness of each experiment. As a result, in such asymptotic regime the "rate" of information that the DM collects remains comparable to those in small sample problems, and therefore, when the DM is experimenting, it does so only because the DM is still unsure of the true hypothesis. This interplay between larger sample sizes and less informative experiments was also recently explored by Naghshvar and Javidi (2013). They also rely on large sample analysis but introduce a multiple hypothesis setting and represent the limited informativeness by scaling the number of hypotheses. Their results are a generalization of Chernoff (1959) in which they suggest adjusted policies and find tight bounds to prove their asymptotic optimality. In the context of multiarmed bandit

(MAB) problems, Wager and Xu (2021) and Fan and Glynn (2021) are two recent arXiv preprints that study a similar type of asymptotic regime and diffusion limits as the ones considered in this paper. In particular, they consider a regime in which the mean rewards of the arms scale as $1/\sqrt{n}$, where n is the number of arm pulls. Wager and Xu (2021) suggest a framework governed by a well-behaved sampling function to implement such approach in the context of sequential experimentation. Fan and Glynn (2021) develop the theory from first principles in the specific context of Thompson sampling. In our two-hypothesis setting, we introduce a general framework to model lack of informativeness. This framework includes, for instance, the case of asymptotically indistinguishable hypothesis as well as settings in which the experiments generate increasingly noisy outcomes. We show that, under our asymptotic regime, the sequential experimentation problem reduces to a diffusion-free boundary problem that we are able to solve and induce approximations for the nonasymptotic regime. Other papers study diffusion models in the context of sequential testing (e.g., Chernoff 1961, Breakwell and Chernoff 1964, Peskir and Shiryaev 2006, Harrison and Sunar 2015) although in our case we make no *Gaussian* assumption regarding the initial process that is being observed. Other examples of sequential analysis papers that rely on diffusion approximations include the work on Bayesian multiarmed bandits by Chang and Lai (1987) and Brezzi and Lai (2002), ranking and selection problems by Chick and Gans (2009) and Chick and Frazier (2012), and also in the context of strategic experimentation with Bolton and Harris (1999), who consider a many-agent, two-armed Bernoulli bandit problem in which agents can learn from the experimentation of other agents (i.e., information as a public good).

Our work also contributes to a growing stream of sequential hypothesis testing problems in the context of best arm identification (BAI) (see Garivier and Kaufmann 2016, Kaufmann et al. 2016, Russo 2020). In our sequential hypothesis testing setup, we interpret each available experiment as an "arm" that, when pulled, generates information on the true hypothesis. A key difference between our model and this literature is that we allow for the possibility that the set of arms available for learning is different from the sets of arms from which the DM chooses a final action. One feature of our model is that the DM learns about the true hypothesis from any pulled arm. This behavior is similar to some BAI settings in which the unknown parameters can affect the reward of multiple correlated arms (see Soare et al. 2014). Moreover, in the illustrative example of Section 6, we assume that an experiment is an assortment of products offered to a customer, and the outcome is the product selected by

that customer. We assume in this example that this selection happens following an MNL model, making this setup similar to an MNL-bandit-like exploration (see the recent work of Agrawal et al. 2019, Oh and Iyengar 2019). Despite some structural difference with BAI and, more generally, MAB literature, we compare in Online Appendix B the numerical performance of some MNL-bandit algorithms—introduced in the literature—with the ones we suggest here.

Finally, we recall that this work naturally belongs to the broad area of reinforcement learning. Our suggested heuristics can be viewed as approximate dynamic programming (ADP) techniques for solving a dynamic learning problem. Such techniques are shown to be effective in managing the curse of dimensionality (see Powell 2016). In this recent review, Powell (2016) divides ADP policies in four categories: myopic cost function approximations, lookahead policies, policy function approximations, and policies based on value function approximations. The latter two are often based on the specific structure of the problem. Indeed, most of the heuristics we suggest (see Section 5) belong to these two categories and are obtained by either reducing carefully the set of policies on which we are optimizing or approximating the value function itself. These approximations are primarily inspired and obtained based on our asymptotic analysis. In our numerical analysis (see Section 7) we also include a lookahead-type policy. Some recent works highlight the effectiveness of simple policies in the context of dynamic learning, such as greedy algorithms (e.g., Bastani et al. 2021) and certainty equivalence (CE; e.g., Keskin and Zeevi 2018). The greedy algorithm behaves well when exploration is expensive although in our case it is free. As for the CE, it is not appropriate in our setting. Indeed, in a Bayesian setting, CE would assume that the current belief is constant moving forward and, hence, would always recommend to stop and never to explore. Having said that, we do show that in our case a simple (static) experimentation policy behaves well and is asymptotically optimal.

Our paper also contributes to the operations literature on sequential testing and demand learning. There is a growing stream of papers in revenue management that focus on the problem of characterizing optimal dynamic pricing strategies when there is incomplete information about consumers' price sensitivity (see Araman and Caldentey 2011, den Boer 2015). In this context, pricing strategies play a dual role. On one hand, they have a direct impact on sales and revenues. On the other, they act as tools for experimentation used by sellers to learn demand characteristics. Optimal pricing strategies are those that balance the so-called exploration–exploitation trade-off between these two roles, for example, Araman and Caldentey

(2009), Besbes and Zeevi (2009), Harrison et al. (2012), den Boer and Zwart (2014), Broder and Rusmevichientong (2012), Gallego and Talebian (2012), and Keskin and Zeevi (2014). Another stream of papers, which is closer to the crowd-voting example that we consider in Section 6, focuses on optimal assortment planning under unknown demand characteristics. In this literature, the decision maker wants to identify a revenue-maximizing assortment of products from a (possibly very large) set of available options. Consumers' preferences over assortments are typically described in the form of a Luce-type choice model—with the MNL being by far the most popular choice—with unknown parameters. In this setting, the DM experiments by displaying different assortments to different consumers over time. Some representative papers in this area include Caro and Gallien (2007), Ulu et al. (2012), Sauré and Zeevi (2013), Agrawal et al. (2019), and Feng et al. (2021). A variant of this line of research is the recent paper of Keskin and Birge (2019), in which the seller faces unknown cost functions that increase with the quality of the products. At each period, the firm selects vertically differentiated products and self-selection pricing mechanisms to learn and maximize its profit over a finite horizon.

Finally, our research also contributes to the recent and growing literature on crowdsourcing and specifically crowd voting. We mention that the recent work of Papanastasiou et al. (2018) tackles the provision of information dissemination in an online setting in which customers' selection of products/services is affected by historical outcomes. On the crowdfunding end, Alaei et al. (2016) suggest a dynamic model of crowdfunding and assess the probability of success of a campaign by introducing the notion of anticipating random walks. On crowd voting, the paper by Marinesi and Girotra (2013) focuses on measuring the information that is acquired from a customer voting system. Using a two-period game-theoretical model, they prove, among other results, that, by offering a sufficiently high discount during the voting phase, crowd-voting systems—used to decide whether to develop the product or not—represent an effective way to elicit information on customers willingness to pay. Our focus, however, through our crowd-voting example, is on how to operationally manage a voting platform that faces a stream of myopic (nonstrategic) consumers. In that regard, our work is close to Feng et al. (2021).

3. Model Description

We consider a DM who must choose an action a in order to maximize a reward function $\mathcal{R}(a, \Theta)$ that depends on both the action a as well as a parameter Θ .

The DM selects the action a from a finite set of available actions \mathcal{A} and does not know the value of Θ that can take one of two possible values $\{\theta_0, \theta_1\}$. Specifically, the DM has incomplete information on Θ and, hence, on the reward function, having a prior that $\Theta = \theta_0$ with probability $\delta \in (0, 1)$.

We assume that the decision maker is risk neutral. If the DM were to make a decision at time $t = 0$, the DM would then select an action that maximizes the DM's expected reward conditional on the DM's prior belief δ . That is, the DM would select an action $a^* \in \mathcal{A}$ that maximizes $\mathbb{E}_\delta[\mathcal{R}(a, \Theta)]$, where $\mathbb{E}_\delta[\cdot]$ is the expectation operator conditional on the prior belief that $\Theta = \theta_0$ with probability δ .⁴ We define the optimal expected reward function

$$\begin{aligned} G(\delta) &:= \max_{a \in \mathcal{A}} \mathbb{E}_\delta[\mathcal{R}(a, \Theta)] \\ &= \max_{a \in \mathcal{A}} \{\delta \mathcal{R}(a, \theta_0) + (1 - \delta) \mathcal{R}(a, \theta_1)\} \end{aligned} \quad (1)$$

and let $\mathcal{A}^*(\delta) \subseteq \mathcal{A}$ be the set of actions at which the maximum reward is achieved. Without loss of optimality, we assume that, for every $a \in \mathcal{A}$, there exists a $\delta \in (0, 1)$ such that $a \in \mathcal{A}^*(\delta)$ (otherwise, some actions are uniformly dominated and can be removed from the set \mathcal{A} of available actions). It is worth noticing that, because \mathcal{A} is a finite set, the function $G(\delta)$ is piecewise linear in δ .

3.1. The Experimentation Process

Instead of selecting immediately an action from the set $\mathcal{A}^*(\delta)$, the DM has the option of postponing this decision in order to experiment and gather additional information about the true value of Θ . The type of experimentation process that we consider is characterized by two key features:

1. The decision maker has at the DM's disposal a finite set \mathcal{E} of experiments. Each experiment $\mathcal{E} \in \mathcal{E}$ has associated a finite set $\mathcal{X}_\mathcal{E}$ of possible outcomes and a likelihood function

$$\mathcal{L}(x, \mathcal{E}) := \frac{Q(x, \mathcal{E}, \theta_1)}{Q(x, \mathcal{E}, \theta_0)}, \quad x \in \mathcal{X}_\mathcal{E},$$

where $Q(x, \mathcal{E}, \theta) := \mathbb{P}_\theta(x | \mathcal{E})$ is the conditional probability of observing outcome $x \in \mathcal{X}_\mathcal{E}$ when the experiment \mathcal{E} is used and $\Theta = \theta$. We assume that every experiment $\mathcal{E} \in \mathcal{E}$ is informative in the sense that there exists $x \in \mathcal{X}_\mathcal{E}$ such that $\mathcal{L}(x, \mathcal{E}) \neq 1$.

2. There exists an exogenous Poisson process N_t , with rate Λ , that determines the time epochs $\{t_i\}_{i \geq 1}$ at which experiments are conducted, where $t_i = \inf\{t \geq 0 : N_t \geq i\}$. As a result, although the decision maker selects the experiment at each experimentation epoch, the DM does not have control over the exact times when these experiments are conducted.⁵

In this setting, a policy is a triplet (π, τ, a_τ) , where π is an experimentation policy that adaptively determines the sequence of experiments $\{\mathcal{E}_{t_1}, \mathcal{E}_{t_2}, \dots\}$ to conduct at the experimentation epochs $\{t_i\}_{i \geq 1}$, τ is a stopping time that defines the duration of the experimentation process, and $a_\tau \in \mathcal{A}$ is the action taken at time τ . Because, at optimality, we have $a_\tau^* \in \mathcal{A}^*(\delta_\tau)$, we simply denote by (π, τ) a generic policy. We also denote by $\{x_{t_1}, x_{t_2}, \dots, x_{t_{N_t}}\}$ the sequence of outcomes of the experiments and by \mathcal{F}_t the history (filtration) generated by the experimentation process up to time t . We denote by \mathbb{T} the set of stopping times with respect to $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$. Also, and using a slight abuse of notation, we denote by \mathcal{E}_t the experiment that is used at time t and by x_t the corresponding outcome. Naturally, we must have $x_t \in \mathcal{X}_{\mathcal{E}_t}$.

By judiciously selecting an experimentation policy π and observing the outcomes of each experiment, the decision maker can gradually learn the true value of Θ over time. In particular, we define the belief process $\delta_t := \mathbb{P}_\delta(\Theta = \theta_0 | \mathcal{F}_t)$ whose evolution is governed by Bayes rule.

Lemma 1 (Belief Process). *Let $\{\mathcal{E}_{t_1}, \mathcal{E}_{t_2}, \dots\}$ be a sequence of experiments and $\{x_{t_1}, x_{t_2}, \dots\}$ be the corresponding sequence of observed outcomes. If the decision maker has a prior belief $\delta = \mathbb{P}_\delta(\Theta = \theta_0)$, then the belief process δ_t evolves as an \mathcal{F}_t -martingale given by*

$$\begin{aligned} \delta_t &= \frac{\delta}{\delta + (1 - \delta)L_t}, \text{ where } L_t \text{ is the likelihood} \\ &\text{-ratio function } L_t := \prod_{i=1}^{N_t} \mathcal{L}(x_{t_i}, \mathcal{E}_{t_i}). \end{aligned} \quad (2)$$

Proof. This and other proofs are relegated to Online Appendix A. \square

3.2. The Optimization Problem

Under some mild assumptions on the likelihood ratios $\mathcal{L}(x, \mathcal{E})$, the belief process converges to zero or one depending on whether $\Theta = \theta_0$ or $\Theta = \theta_1$, respectively. Hence, an infinitely patient decision maker eventually learns the true value of Θ . However, by running a long experimentation process, the decision maker is also delaying the time when the final decision is made. If we assume that, ceteris paribus, the decision maker prefers to collect these rewards as early as possible, then the DM faces a trade-off between learning the true value of Θ (exploration) and collecting the reward $\mathcal{R}(a, \Theta)$ (exploitation). To model this trade-off, we assume that the decision maker's objective is to maximize the expected discounted reward that the DM will collect at the time a final

decision is made. That is, the DM is interested in solving the following optimal stopping time problem:

$$\Pi(\delta) := \sup_{(\pi, \tau)} \mathbb{E}_\delta[e^{-r\tau} G(\delta_\tau)], \quad (3)$$

where r is the decision maker's discount factor. We tackle the solution of (3) using dynamic programming. To this end, we find it convenient to express the dynamic evolution of the belief process δ_t in Equation (2) using the following stochastic differential equation (SDE) representation.

Lemma 2. *The belief process in (2) admits the SDE representation*

$$d\delta_t = \eta(\delta_{t-}, x_t, \mathcal{E}_{t-}) dN_t, \quad \text{where}$$

$$\eta(\delta, x, \mathcal{E}) := (1 - \delta)\delta \left(\frac{1 - \mathcal{L}(x, \mathcal{E})}{\delta + (1 - \delta)\mathcal{L}(x, \mathcal{E})} \right).$$

In the statement of the previous lemma, the left-limit notation \mathcal{E}_{t-} (δ_{t-}) stands for the experiment (belief) that is chosen (observed) right before a jump of N_t at time t . The factor $\eta(\delta, x, \mathcal{E})$ is the size of the jump of the belief process (i.e., the “amount” of learning) if an experiment \mathcal{E} is chosen that produces an outcome x when the belief process (just before the experiment) is equal to δ .

Equipped with Lemma 2, we formulate the decision maker's problem as a Markov decision problem and, without loss of optimality, restrict our attention to the class of deterministic Markovian policies (e.g., Blackwell 1965 and section 4.4 in Puterman 2005). In particular, the experimentation policy π maps each value of the belief δ to an experiment $\pi(\delta) \in \mathcal{E}$, and the stopping time τ is a hitting time of the belief process on some *intervention* set \mathcal{I} . We interchangeably use τ and \mathcal{I} depending on the context. In the following definition, $\mathcal{M}(\mathcal{E})$ is the set of measurable functions from $[0, 1]$ to \mathcal{E} and \mathcal{B} is the set of Borel sets in $[0, 1]$.

Definition 1 (Deterministic Markovian Policy). A deterministic Markovian policy corresponds to a pair $(\pi, \mathcal{I}) \in \mathcal{M}(\mathcal{E}) \times \mathcal{B}$. For all $\delta \notin \mathcal{I}$, the DM displays experiment $\pi(\delta) \in \mathcal{E}$. On the other hand, for $\delta \in \mathcal{I}$ the decision maker chooses to stop the experimentation process and implements an optimal action $a \in \mathcal{A}^*(\delta)$.

Putting all the pieces together, the decision maker's optimization in (3) can be rewritten as the following optimal control problem:

$$\Pi(\delta) := \sup_{(\pi, \mathcal{I})} \mathbb{E}_\delta[e^{-r\tau} G(\delta_\tau)] \quad (4)$$

$$\text{subject to: } d\delta_t = \eta(\delta_{t-}, x_t, \pi(\delta_{t-})) dN_t, \quad \delta_0 = \delta,$$

$$\text{and } \tau = \inf \{t > 0 : \delta_t \in \mathcal{I}\}.$$

The following proposition proves useful in various places in the analysis that follows.

Proposition 1. *The functions $G(\delta)$ and $\Pi(\delta)$ are both convex in $\delta \in [0, 1]$.*

In Online Appendix B, we discuss how to exploit the convexity of $\Pi(\delta)$ to simplify the optimization problem by eliminating some experiments that are dominated in a *convex order dominance* sense.⁶

To solve the control problem in (4), we can express its optimality conditions in the form of the following Hamilton–Jacobi–Bellman (HJB) equation:

$$0 = \max \left\{ G(\delta) - \Pi(\delta), \Lambda \max_{\mathcal{E} \in \mathcal{E}} \{ \mathbb{E}_\delta[\Pi(\delta + \eta(\delta, x, \mathcal{E})) - \Pi(\delta)] \} - r\Pi(\delta) \right\}, \quad (5)$$

with border conditions $\Pi(0) = G(0)$ and $\Pi(1) = G(1)$ because both $\delta = 0$ and $\delta = 1$ are absorbing belief states (see Lemma 1). By solving the inner maximization, we can compute an optimal experimentation policy $\mathcal{E}^*(\delta)$, that is,

$$\mathcal{E}^*(\delta) \in \arg \max_{\mathcal{E} \in \mathcal{E}} \{ \mathbb{E}_\delta[\Pi(\delta + \eta(\delta, x, \mathcal{E}))] \}. \quad (6)$$

The HJB equation in (5) leads to a tractable computational approach to solve the decision maker's problem. For instance, we can implement the *value iteration* algorithm

$$\Pi_0(\delta) = G(\delta) \quad \text{and}$$

$$\Pi_{l+1}(\delta) = \max \left\{ G(\delta), \frac{\Lambda}{\Lambda + r} \max_{\mathcal{E} \in \mathcal{E}} \{ \mathbb{E}_\delta[\Pi_l(\delta + \eta(\delta, x, \mathcal{E}))] \} \right\}, \quad (7)$$

which defines a sequence of continuous functions $\{\Pi_l(\delta) : l \geq 0\}$ that are monotonically increasing in l and converge uniformly to a limit $\Pi(\delta) = \lim_{l \rightarrow \infty} \Pi_l(\delta)$ that satisfies the HJB equation in (5), (see the proof of Proposition 1 for details).

Despite its computational simplicity, the HJB Equation (5) is not particularly malleable for the purpose of analysis and to derive structural results about an optimal solution and its properties. For this reason, in the next sections, we tackle the decision maker's optimization problem using a diffusion approximation that preserves the same trade-offs as in the original formulation but provides a more transparent representation of the problem and its optimal solution.

We end this section with a numerical example that illustrates the value iteration method used and highlights some feature of an optimal solution.

Example 1. Suppose the decision maker has four alternative actions from which to choose (i.e., $|\mathcal{A}| = 4$) with corresponding payoffs $\mathcal{R}_1(\delta) = 6 - 30\delta$, $\mathcal{R}_2(\delta) = 4 - 5\delta$, $\mathcal{R}_3(\delta) = 3\delta$, and $\mathcal{R}_4(\delta) = -20 + 25\delta$. There are nine possible experiments that the DM can use (i.e., $|\mathcal{E}| = 9$), and each experiment produces a binary outcome, that is, $\mathcal{X}_\mathcal{E} = \{0, 1\}$ for all $\mathcal{E} \in \mathcal{E}$. Table 1 specifies the probability $Q(0, \mathcal{E}, \Theta)$ for each of the nine

Table 1. Probabilities of Observing Outcome Zero for Each of the Nine Experiments as a Function of the Value of Θ

	$Q(0, \mathcal{E}, \Theta)$								
Experiment	1	2	3	4	5	6	7	8	9
$\Theta = \theta_0$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\Theta = \theta_1$	0.03	0.04	0.09	0.16	0.25	0.36	0.49	0.68	0.86

experiments for $\Theta = \theta_0$ and $\Theta = \theta_1$. Finally, we let $\Lambda = 8$ and $r = 0.5$.

Figure 1 depicts the numerically computed solution using the value iteration in (7) after 200 iterations. The left panel shows the value function $\Pi(\delta)$, and the right panel shows the optimal experiment $\mathcal{E}^*(\delta)$. We use the convention $\mathcal{E}^*(\delta) = 0$ for those values of δ at which $\Pi(\delta) = G(\delta)$ and no experimentation is used.

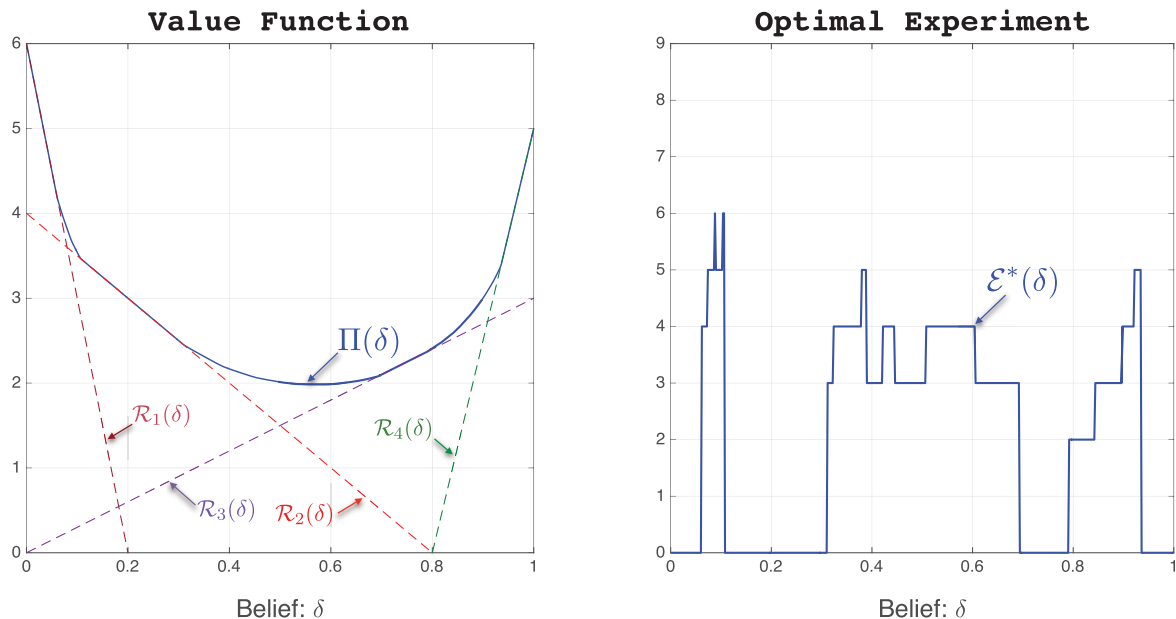
As we can see, in an optimal solution, the belief space is partitioned into a collection of intervals that define the regions in which experimentation is used or not used. For example, in the interval $\delta \in [0.1, 0.31]$, the decision maker does not use any experimentation and selects immediately (at time $t = 0$) an action in $\mathcal{A}^*(\delta)$ that maximizes the DM’s expected reward (in this case $\mathcal{A}^*(\delta) = \{2\}$). On the other hand, in the interval $\delta \in (0.31, 0.69)$, the DM wants to experiment. In this case, the interval $(0.31, 0.69)$ is further partitioned into a collection of subintervals in which a specific experiment is selected. For instance, for $\delta \in (0.45, 0.51)$, the decision maker uses Experiment 3, and for $\delta \in (0.51, 0.61)$, the DM uses Experiment 4.

We note that experimentation occurs around those values of δ for which the payoff function $G(\delta) = \max_i \{\mathcal{R}_i(\delta)\}$ has a kink, that is, for which two payoff functions intersect. Intuitively, in these regions, a small change in the value of δ can lead to a discrete change in the optimal action to select, and so the DM has locally more incentive to experiment and learn in these regions.

4. Asymptotic Approximation

In this section, we specialize the problem described in the previous section to a particular class of instances in which (i) experiments are conducted at “high frequency” although (ii) the “informativeness” of each experiment is low. There are many natural and practical situations in which the decision maker has access to a large number of experiments but the informativeness of each individual one is low. For instance, online experiments are becoming quite common in the business world, and each experiment is often linked to one visitor who is offered a set of choices from which to select. Such common setup generates a large volume of experiments in a relatively short time period. However, one of the major issues faced by the experimenter is the relevance and veracity of the data generated (we refer the reader to the section “Beware of Low-Quality Data” in Kohavi and Thomke 2017). In some cases, the heterogeneity of the experimentees in online experimentation can generate very noisy data. Moreover, the hypotheses being tested can be marginally different, making the task of distinguishing them

Figure 1. (Color online) Numerically Computed Solution



Notes. The left panel depicts the value functions $\Pi(\delta)$. The right panel depicts the optimal experiment $\mathcal{E}^*(\delta)$. Data: $\mathcal{R}_1(\delta) = 6 - 30\delta$, $\mathcal{R}_2(\delta) = 4 - 5\delta$, $\mathcal{R}_3(\delta) = 3\delta$, $\mathcal{R}_4(\delta) = -20 + 25\delta$, $r = 0.5$, and $\Lambda = 8$.

harder. Both settings are typical and are examples of how little informative online experiments can be. Our illustrative example in Section 6 builds on these ideas. The notion of limited informativeness of experiments is also present in other settings. In their recent work, Lewis and Rao (2015) show how difficult it is to prove the return on investment of advertising campaigns. The paper notes specifically that informative advertising experiments can require more than 10 million person-weeks, which reflects exactly the tension of our regime between large sample size and little informativeness. Clinical trials are another major area that suffers from serious data error and lack of accuracy (see Nahm 2012) and, as a result, would require large sample sizes.

4.1. Diffusion Formulation

To formalize the notion of a *high frequency versus low informativeness* regime of experimentation, we consider a sequence of instances of the problem indexed by a nonnegative integer k in such a way that, as k grows large, both the number of experiments conducted per unit of time becomes high and the amount of information generated from each individual experiment goes to zero. Under our proposed scaling, the magnitude of the jumps of δ_t as well as the time between two experiments converge to zero, resulting in a belief process that converges weakly to a diffusion process.

We let $Q^k(x, \mathcal{E}, \theta)$ be the conditional probability of observing outcome $x \in \mathcal{X}_{\mathcal{E}}$ when experiment $\mathcal{E} \in \mathcal{E}$ is conducted conditional on $\Theta = \theta$ for the k^{th} instance of the problem. To capture the notion of low informativeness of an experiment, we impose the following requirement on the sequence $\{Q^k(x, \mathcal{E}, \theta)\}$.

Assumption 1 (Low-Informativeness Regime). *For each $\mathcal{E} \in \mathcal{E}$, there exists a probability distribution $Q(\cdot, \mathcal{E})$ such that, for $\theta \in \{\theta_0, \theta_1\}$,*

$$\sqrt{k} \left(\frac{Q^k(x, \mathcal{E}, \theta)}{Q(x, \mathcal{E})} - 1 \right) \rightarrow \alpha(x, \mathcal{E}, \theta), \quad (8)$$

where $\alpha(x, \mathcal{E}, \theta)$ satisfies

$$\sum_{x \in \mathcal{X}_{\mathcal{E}}} \alpha(x, \mathcal{E}, \theta) Q(x, \mathcal{E}) = 0.$$

Intuitively, the asymptotic scaling in Assumption 1 has the following property: as $k \rightarrow \infty$, the likelihood function $\mathcal{L}^k(x, \mathcal{E}) = Q^k(x, \mathcal{E}, \theta_1)/Q^k(x, \mathcal{E}, \theta_0)$ converges to one for every $x \in \mathcal{X}_{\mathcal{E}}$, and as a result, the jumps $\eta^k(\delta, x, \mathcal{E})$ of δ_t (see Lemma 2) converge to zero. In other words, in this asymptotic regime, the outcomes of an experiment become less and less informative as k grows large.

On its own, the scaling in Equation (8) would lead to a trivial limit in which δ_t remains constant over time. To counterbalance the fact that individual experiments become less informative under (8), we also scale up the arrival rate of N_t in a way that the amount

of information collected by the experimentation process per unit of time remains comparable to the one in the original unscaled system. Specifically, let N_t^k denote the Poisson process that determines the experimentation epochs for the k^{th} instance.

Assumption 2 (High-Frequency Regime). *Let Λ^k be the intensity of N_t^k . Then, Λ^k satisfies*

$$\Lambda^k = k \Lambda, \quad (9)$$

for some fixed constant $\Lambda > 0$.

Our objective at this point is to suggest a diffusion approximation of the general formulation Problem (4). For that, we combine the parameter scalings in (8) and (9) to obtain a well-defined diffusion limit for the belief process, δ_t .⁷ We derive this limit over the class of continuous randomized Markovian policies defined as follows and show that this class contains an ε -optimal policy for any $\varepsilon > 0$ (see Proposition 3).

In the following definition, $\Delta(\mathcal{E})$ is the set of probability distributions on the collection of possible experiments in \mathcal{E} and $\mathcal{M}_c(\Delta(\mathcal{E}))$ is the set of continuous measurable functions from $[0, 1]$ to $\Delta(\mathcal{E})$. For a randomized experimentation policy $\pi \in \mathcal{M}_c(\Delta(\mathcal{E}))$, we let $\pi(\delta, \mathcal{E})$ denote the probability of selecting experiment \mathcal{E} , which is continuous in δ for all $\mathcal{E} \in \mathcal{E}$.

Definition 2 (Continuous Randomized Markovian Policy). A continuous randomized Markovian policy is a pair $(\pi, \mathcal{I}) \in \mathcal{M}_c(\Delta(\mathcal{E})) \times \mathcal{B}$. For all $\delta \in \mathcal{I}$, the DM displays experiment $\mathcal{E} \in \mathcal{E}$ with probability $\pi(\delta, \mathcal{E})$. On the other hand, for $\delta \in \mathcal{I}$ the decision maker chooses to stop the experimentation process and implements an optimal action $a \in \mathcal{A}^*(\delta)$.

Next, we move to state our limiting result for the belief process under the scalings given in (8) and (9).

Proposition 2. *Consider a fixed experimentation policy $\pi \in \mathcal{M}_c(\Delta(\mathcal{E}))$ and let δ_t^k be the belief process induced by π for instance k under the scaling in Equations (8) and (9). Then, we have that $\delta_t^k \Rightarrow \tilde{\delta}_t$ as $k \rightarrow \infty$, where $\tilde{\delta}_t$ is a diffusion process solution of the SDE*

$$d\tilde{\delta}_t = \tilde{\sigma}(\tilde{\delta}_t, \pi) \tilde{\delta}_t (1 - \tilde{\delta}_t) dW_t,$$

where W_t is a Wiener process and

$$\tilde{\sigma}^2(\delta, \pi) := \Lambda \sum_{\mathcal{E} \in \mathcal{E}} \sum_{x \in \mathcal{X}_{\mathcal{E}}} \pi(\delta, \mathcal{E}) (\alpha(x, \mathcal{E}, \theta_1) - \alpha(x, \mathcal{E}, \theta_0))^2 Q(x, \mathcal{E}). \quad (10)$$

Remark 1. Throughout the paper, we use tildes (\sim) to denote quantities that are related to the asymptotic approximation. \diamond

The next result shows that restricting our attention to continuous randomized policies is without a

significant loss of optimality in the sense of the L^1 norm.

Proposition 3. *Let $(\pi, \mathcal{I}) \in \mathcal{M}(\mathcal{E}) \times \mathcal{B}$ be an optimal Markovian policy with a corresponding value function $\Pi(\delta)$. For any $\varepsilon > 0$, there exists a continuous randomized policy (π_c, \mathcal{I}_c) in $\mathcal{M}_c(\mathcal{E}) \times \mathcal{B}$ with expected payoff function $\tilde{\Pi}(\delta)$ such that*

$$\|\Pi - \tilde{\Pi}\|_1 < \varepsilon,$$

where $\|\cdot\|_1$ is the L^1 norm in $[0, 1]$.

In sum, and in light of Propositions 2 and 3, we suggest the following diffusion-asymptotic approximation of the decision maker’s problem given in (4):

$$\begin{aligned} \tilde{\Pi}(\delta) &= \sup_{(\pi, \mathcal{I}) \in \mathcal{M}_c(\mathcal{E}) \times \mathcal{B}} \mathbb{E}_\delta[e^{-r\tau} G(\tilde{\delta}_\tau)] \quad \text{s.t.} \\ d\tilde{\delta}_t &= \tilde{\sigma}(\delta_t, \pi) \tilde{\delta}_t (1 - \tilde{\delta}_t) dW_t \quad \text{and} \\ \tau &= \inf\{t > 0 : \tilde{\delta}_t \in \mathcal{I}\}. \end{aligned} \tag{11}$$

We can view Problem (11) as having two decision variables, namely, the experimentation policy π and the intervention region \mathcal{I} . Interestingly, it turns out that we can decouple the optimization of these two decisions; in particular, we can solve for the optimal experimentation π^* without computing explicitly \mathcal{I}^* . Surprisingly, this implies that the choice of an optimal experiment is independent of the intervention region and, thus, of how long the decision maker decides to run the experimentation process. We formalize this observation in the following section.

4.2. Asymptotically Optimal Experimentation Policy

From the diffusion approximation in Equation (11), one can easily see that the impact an experimentation policy π has on the decision maker’s optimization problem is channeled only through the volatility of the belief process $\tilde{\sigma}(\delta, \pi)$. We use this fact to derive a rather simple solution to the problem of selecting an asymptotically optimal policy π^A . (The superscript “A” is mnemonic of “asymptotic”). To this end, let us define the mapping

$$T_t^\pi := \int_0^t \frac{1}{\tilde{\sigma}^2(\delta_s, \pi)} ds,$$

which acts as a random time change in the following proposition.

Proposition 4. *The optimization problem in (11) is equivalent to*

$$\begin{aligned} \tilde{\Pi}(\delta) &= \sup_{(\pi, \mathcal{I}) \in \mathcal{M}_c(\mathcal{E}) \times \mathcal{B}} \mathbb{E}_\delta[e^{-rT_\tau^\pi} G(\tilde{\delta}_\tau)] \quad \text{s.t.} \\ d\tilde{\delta}_t &= \tilde{\delta}_t (1 - \tilde{\delta}_t) dW_t \quad \text{and} \quad \tau = \inf\{t > 0 : \tilde{\delta}_t \in \mathcal{I}\}. \end{aligned}$$

The previous result provides an alternative interpretation of the effect an experimentation policy π has on

the decision maker’s performance. According to Proposition 4, a policy π impacts only the discount factor rT_τ^π that the decision maker uses to penalize the time value of money. The following corollary follows directly from this observation.

Corollary 1 (Maximum Volatility). *For any stopping set $\mathcal{I} \in \mathcal{B}$, an optimal asymptotic experimentation policy π^A minimizes the modified discount factor rT_τ^π pathwise or, equivalently, maximizes pointwise the belief process’s volatility $\tilde{\sigma}^2(\delta, \pi)$. Thus, from (10), we conclude that we can select π^A to be a static experimentation policy, namely, $\pi^A(\delta, \mathcal{E}) = 11(\mathcal{E} = \tilde{\mathcal{E}}^A)$ for all $\delta \in \mathcal{I}^c$, where \mathcal{E}^A is given by*

$$\tilde{\mathcal{E}}^A = \arg \max_{\mathcal{E} \in \mathcal{E}} \left\{ \sum_{x \in \mathcal{X}_\mathcal{E}} (\alpha(x, \mathcal{E}, \theta_1) - \alpha(x, \mathcal{E}, \theta_0))^2 \mathcal{Q}(x, \mathcal{E}) \right\}.$$

(If there are multiple experiments that maximize the expression inside the brackets, then we can select any static experimentation policy that uses an experiment $\tilde{\mathcal{E}}^A$ from the argmax set.)

A few remarks about this result are in order. First, Corollary 1 confirms our previous claim that an optimal experimentation policy is independent of the choice of the stopping set \mathcal{I} , and so we can effectively decouple the problem of determining an optimal experimentation strategy and that of when to stop experimenting. We also note that an optimal static experimentation strategy is continuous in δ , and so we can invoke the weak convergence in Proposition 2 directly to $\pi^A(\delta, \tilde{\mathcal{E}}^A)$.

Example 2 (Example 1 Revisited). To illustrate the result in Corollary 1, let us revisit the instance in Example 1 in the context of the asymptotic regime. To this end, suppose the probabilities $Q^k(0, \mathcal{E}, \theta)$ for the k^{th} instance of the problem are equal to

$$Q^k(0, \mathcal{E}, \theta) = Q(0, \mathcal{E}) \left(1 + \frac{\alpha(0, \mathcal{E}, \theta)}{\sqrt{k}} \right),$$

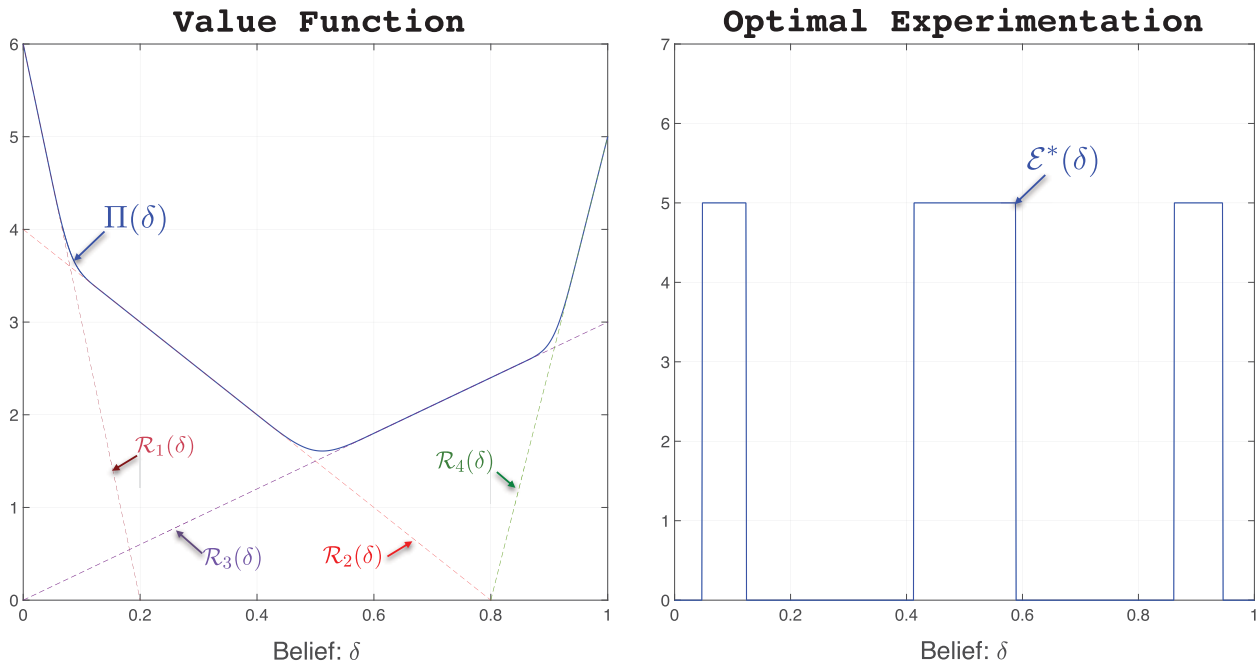
where $Q(0, \mathcal{E})$ and $\alpha(0, \mathcal{E}, \theta)$ are given in Table 2. Note that we are not including Experiments 1, 8, and 9 because they are dominated (see Online Appendix B). Also, the original probabilities in Table 1 correspond to the case $k = 1$. Figure 2 mimics Figure 1 but for $k = 10,000$.

Consistent with the result in Corollary 1, for k sufficiently large, the optimal experimentation strategy

Table 2. The Values of $Q, \alpha(\hat{c}, \hat{c}, \theta_0)$ and $\alpha(\hat{c}, \hat{c}, \theta_1)$ for an Outcome “0” and for Experiments 2, 3, 4, 5, 6, and 7

Experiment	2	3	4	5	6	7
$Q(0, \mathcal{E})$	0.2	0.3	0.4	0.5	0.6	0.7
$\alpha(0, \mathcal{E}, \theta_0)$	0.0	0.0	0.0	0.0	0.0	0.0
$\alpha(0, \mathcal{E}, \theta_1)$	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3

Figure 2. (Color online) Numerically Computed Solution Under the Asymptotic Regime



Note. Data: $\mathcal{R}_1(\delta) = 6 - 30\delta$, $\mathcal{R}_2(\delta) = 4 - 5\delta$, $\mathcal{R}_3(\delta) = 3\delta$, $\mathcal{R}_4(\delta) = -20 + 25\delta$, $r = 0.5$, $\Lambda = 8k$, and $k = 10,000$.

$\mathcal{E}^A(\delta)$ consists of a single experiment independent of δ , which in this case corresponds to Experiment 5. One can check that Experiment 5 maximizes the instantaneous volatility of the belief process. \diamond

4.3. Optimal Stopping of Experimentation

Let us now turn to the problem of determining the optimal intervention region in the asymptotic regime under consideration. In what follows, we assume that an optimal experimentation policy has been selected based on Corollary 1. That is, we focus on solving Problem (11) given the optimal experimentation policy $\pi^A(\delta, \mathcal{E}) = \mathbb{1}(\mathcal{E} = \tilde{\mathcal{E}}^A)$. We find it convenient to rewrite this problem using the following optimal stopping time formulation:

$$\begin{aligned} \tilde{\mathcal{G}}(\delta) &:= \sup_{\tau \in \mathbb{T}} \mathbb{E}_\delta \left[e^{-r\tau} G(\tilde{\delta}_\tau) \right] \quad \text{subject to} \\ d\tilde{\delta}_t &= \tilde{\sigma} \tilde{\delta}_t (1 - \tilde{\delta}_t) dW_t, \quad \tilde{\delta}_0 = \delta. \end{aligned} \tag{12}$$

For notational convenience, throughout this section, we suppress the dependence of $\tilde{\mathcal{G}}$ and $\tilde{\sigma}$ on the display set $\tilde{\mathcal{E}}^A$ because it remains fixed.

We approach the problem in two steps. First, we derive optimality conditions in the form of a set of partial differential inequalities that characterize the optimal stopping time. Then, we use these inequalities to characterize an optimal solution and the corresponding payoff.

4.3.1. Quasi-Variational Inequalities (QVI). Let $\mathcal{C}^k[0, 1]$ denote the set of real-valued continuous functions on $[0, 1]$ having derivatives of order $k \geq 0$. We define also the set

$$\begin{aligned} \tilde{\mathcal{C}}^2 &:= \{f \in \mathcal{C}^1[0, 1] : \text{there exists a finite set } N_f \subseteq [0, 1] \\ &\text{such that } f''(\delta) \text{ exists } \forall \delta \in [0, 1] \setminus N_f\}. \end{aligned} \tag{13}$$

(Note that the set N_f depends on the specific function f .) We also define the operator \mathcal{H} on $\tilde{\mathcal{C}}^2$ as follows:

$$\mathcal{H}f(\delta) := \frac{1}{2} \tilde{\sigma} \delta (1 - \delta) f''(\delta) - r f(\delta), \quad \text{for all } \delta \in [0, 1] \setminus N_f. \tag{14}$$

Definition 3 (QVI). The function $f \in \tilde{\mathcal{C}}^2$ satisfies the quasi-variational inequalities for the optimization Problem (12) if, for all $\delta \in [0, 1] \setminus N_f$,

$$\begin{aligned} f(\delta) - G(\delta) &\geq 0 \\ \mathcal{H}f(\delta) &\leq 0 \\ (f(\delta) - G(\delta)) \mathcal{H}f(\delta) &= 0. \quad \square \end{aligned} \tag{15}$$

As one might expect, a solution to these QVI conditions partitions the interval $[0, 1]$ into two regions: a *continuation* region in which the firm's optimal strategy is to keep experimenting and an *intervention* region in which stopping the experimentation process is optimal.

Continuation: $\mathcal{C} := \{\delta \in [0, 1] : f(\delta) > G(\delta) \text{ and } \mathcal{H}f(\delta) = 0\}$,

Intervention: $\mathcal{I} := \{\delta \in [0, 1] : f(\delta) = G(\delta) \text{ and } \mathcal{H}f(\delta) \leq 0\}$.

For every solution of the QVI conditions, we can associate a control $\tau \in \mathbb{T}$.

Definition 4. Let $f \in \widehat{\mathcal{C}}^2$ be a solution of the QVI conditions in (15). We define the control τ as follows:

$$\tau = \inf \{t > 0 : f(\tilde{\delta}_t) = G(\tilde{\delta}_t)\}$$

and refer to it as the QVI-control associated to f .

We are now ready to formalize the verification theorem that provides the connection between the QVI conditions and the original optimization problem in (12).

Theorem 1 (Verification). Let $f \in \widehat{\mathcal{C}}^2$ be a solution of the QVI in (15). Then,

$$f(\delta) \geq \tilde{G}(\delta) \text{ for every } \delta \in [0, 1].$$

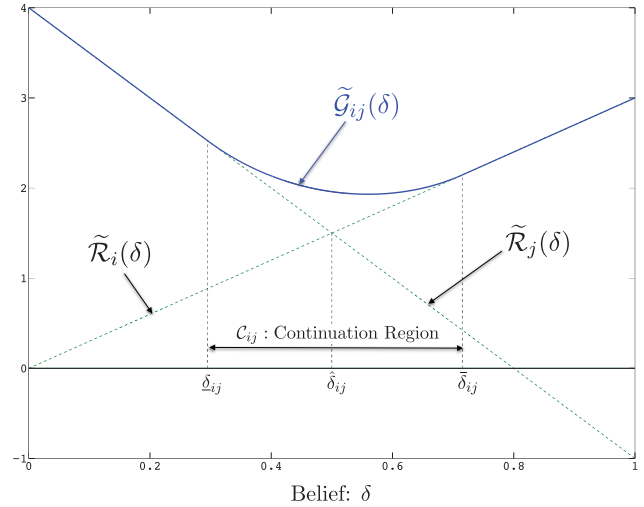
In addition, if there exists a QVI-control τ associated with f such that $\mathbb{E}[\tau] < \infty$, then τ is optimal and $f(\delta) = \tilde{G}(\delta)$.

This verification theorem reduces the problem of determining the value function $\tilde{G}(\delta)$ to that of solving the QVI equations defined earlier. In order to find a solution, we take full advantage of the fact that the payoff function $G(\delta)$ is a piecewise linear continuous function of $\delta \in [0, 1]$ (see Equation (1)). Moreover, an important building block in our methodology is the solution to a special case in which $G(\delta)$ has only two linear pieces, that is, the set \mathcal{A} includes only two actions. We focus on this simpler case first and then show how to leverage this solution and extend it to the general case in which \mathcal{A} includes an arbitrary number of actions.

4.3.2. Special Case: $|\mathcal{A}| = 2$. Suppose the set of actions is given by $\mathcal{A} = \{a_i, a_j\}$ for two distinctive actions, a_i and a_j . Let us denote by $\tilde{G}_{ij}(\delta) = \max\{\tilde{\mathcal{R}}_i(\delta), \tilde{\mathcal{R}}_j(\delta)\}$, where $\tilde{\mathcal{R}}_n(\delta) = \mathbb{E}_\delta[\mathcal{R}(a_n, \Theta)]$ for $n = i, j$ (see Equation (1)). Without loss of generality, we assume that $\tilde{G}_{ij}(\delta) \geq 0$ for all $\delta \in [0, 1]$. Let us denote by $\hat{\delta}_{ij}$ the value of the belief at which $\tilde{\mathcal{R}}_i(\hat{\delta}_{ij}) = \tilde{\mathcal{R}}_j(\hat{\delta}_{ij})$. (Recall that we have assumed there is no action in the set \mathcal{A} that is uniformly dominated, and this assumption guarantees the existence of $\hat{\delta}_{ij} \in [0, 1]$.)

To solve the QVI conditions in this special case we take an “educated guess” approach and assume that the continuation region \mathcal{C}_{ij} is given by an interval $[\underline{\delta}_{ij}, \bar{\delta}_{ij}]$ for two thresholds $0 \leq \underline{\delta}_{ij} \leq \bar{\delta}_{ij} \leq 1$. Furthermore, we assume the intuitive fact that $\hat{\delta}_{ij} \in [\underline{\delta}_{ij}, \bar{\delta}_{ij}]$. To illustrate, consider the example in Figure 3 that depicts the value function $\tilde{G}_{ij}(\delta)$ as well as the payoff functions $\tilde{\mathcal{R}}_i(\delta)$ and $\tilde{\mathcal{R}}_j(\delta)$ for products i and j .

Figure 3. (Color online) Example of the Value Function $\tilde{G}_{ij}(\delta)$ and Continuation Region \mathcal{C}_{ij} for the Case in Which \mathcal{A} Includes Two Actions



Note. Data: $\tilde{\mathcal{R}}_i(\delta) = 3\delta$, $\tilde{\mathcal{R}}_j(\delta) = 4 - 5\delta$, $r = 1$, and $\tilde{\sigma} = 2$.

By definition, in the interior of the continuation region, we have that $\tilde{G}_{ij}(\delta) < \tilde{G}_{ij}(\delta)$. Hence, according to the third QVI condition, in this region, the value function must satisfy $\mathcal{H}\tilde{G}_{ij}(\delta) = 0$. This is a second-order differential equation

$$\frac{(\tilde{\sigma}\delta(1-\delta))^2}{2} \tilde{G}_{ij}''(\delta) - r\tilde{G}_{ij}(\delta) = 0,$$

whose general solution is given by

$$\tilde{G}_{ij}(\delta) = C_{ij}^0 \frac{(1-\delta)^\gamma}{\delta^{\gamma-1}} + C_{ij}^1 \frac{\delta^\gamma}{(1-\delta)^{\gamma-1}}, \tag{16}$$

$$\text{where } \gamma := \frac{1 + \sqrt{1 + 8r/\tilde{\sigma}^2}}{2},$$

and C_{ij}^0 and C_{ij}^1 are two constants of integration.

To complete our proposed characterization of the value function, we need to determine the constants of integration as well as the two thresholds $\underline{\delta}_{ij}$ and $\bar{\delta}_{ij}$ that define the continuation region. To do that, we impose the so-called value-matching and smooth-pasting conditions that regulate the behavior of the value function at the boundaries between the intervention and continuation regions. Specifically, we impose the conditions

$$\tilde{G}_{ij}(\delta) = \tilde{\mathcal{R}}_i(\delta) \quad \text{and} \quad \tilde{G}'_{ij}(\delta) = \tilde{\mathcal{R}}'_i(\delta) \quad \text{for } \delta = \underline{\delta}_{ij}, \bar{\delta}_{ij}. \tag{17}$$

We formalize our previous discussion in the next proposition.

Proposition 5. Let $\gamma = (1 + \sqrt{1 + 8r/\bar{\sigma}^2})/2$. If $\mathcal{A} = \{a_i, a_j\}$, then the QVI conditions admit a solution that we denote by $\tilde{\mathcal{G}}_{ij}(\cdot)$ given as follows:

$$\tilde{\mathcal{G}}_{ij}(\delta) = \begin{cases} \tilde{\mathcal{G}}_{ij}(\delta) & \text{if } 0 \leq \delta \leq \underline{\delta}_{ij} \\ C_{ij}^0(1-\delta)^\gamma \delta^{1-\gamma} + C_{ij}^1(1-\delta)^{1-\gamma} \delta^\gamma & \text{if } \underline{\delta}_{ij} \leq \delta \leq \bar{\delta}_{ij} \\ \tilde{\mathcal{G}}_{ij}(\delta) & \text{if } \bar{\delta}_{ij} \leq \delta \leq 1. \end{cases} \quad (18)$$

Here, $\underline{\delta}_{ij}, \bar{\delta}_{ij} \in (0, 1)$ and C_{ij}^0 and C_{ij}^1 are positive constants all determined by imposing the value-matching and smooth-pasting conditions in (17). The function $\tilde{\mathcal{G}}_{ij}(\delta)$ is convex and in $\tilde{\mathcal{C}}^2$.

The verification theorem guarantees that the solution expressed in Proposition 5 is such that $\tilde{\mathcal{G}}_{ij} = \tilde{\mathcal{G}}$ when $\mathcal{A} = \{i, j\}$. In terms of implementation, this solution corresponds to the following policy.

Asymptotically optimal intervention policy: Suppose the initial belief lies in the interior of the continuation region C_{ij} , that is, $\delta \in (\underline{\delta}_{ij}, \bar{\delta}_{ij})$. In this case, the decision maker runs an experimentation process and keeps it running as long as $\delta_t \in (\underline{\delta}_{ij}, \bar{\delta}_{ij})$. As soon as the belief process δ_t hits one of the two thresholds $\underline{\delta}_{ij}$ or $\bar{\delta}_{ij}$, then the experimentation process stops and the decision maker selects the action that maximizes $\tilde{\mathcal{G}}_{ij}$ at that time. On the other hand, if the initial belief δ is not in the interior of the continuation region, then no experimentation is needed, and the decision maker selects the action that maximizes $\tilde{\mathcal{G}}_{ij}(\delta)$ at time 0. \diamond

The simple representation of the value function in Proposition 5 is due to the diffusion approximation obtained in this asymptotic regime. Moreover, this same diffusion approximation allows one to use some standard results for one-dimensional diffusion processes (e.g., section 5.5 in Karatzas and Shreve 1991) to analyze its optimal solution, for instance, in those cases in which $\delta \in (\underline{\delta}_{ij}, \bar{\delta}_{ij})$ experimentation should be conducted and its duration corresponds to the first exit time of δ_t from the interval $(\underline{\delta}_{ij}, \bar{\delta}_{ij})$. The following corollary characterizes the expected duration of this experimentation phase as well as the likelihood that action a_i or a_j is eventually selected.

Corollary 2. Suppose $\delta \in (\underline{\delta}_{ij}, \bar{\delta}_{ij})$ and let $\tau^* = \inf\{t > 0 : \delta_t \notin (\underline{\delta}_{ij}, \bar{\delta}_{ij})\}$ and $\bar{p}(\delta) = \mathbb{P}(\delta_{\tau^*} = \bar{\delta}_{ij} \mid \delta_0 = \delta)$. Then,

$$\bar{p}(\delta) = \frac{\delta - \underline{\delta}_{ij}}{\bar{\delta}_{ij} - \underline{\delta}_{ij}} \quad \text{and}$$

$$\mathbb{E}[\tau^*] = \bar{p}(\delta) \mathcal{T}(\bar{\delta}_{ij}) + (1 - \bar{p}(\delta)) \mathcal{T}(\underline{\delta}_{ij}) - \mathcal{T}(\delta),$$

where $\mathcal{T}(\delta)$ is the function

$$\mathcal{T}(\delta) := \frac{2}{\bar{\sigma}^2} (2\delta - 1) \ln\left(\frac{\delta}{1-\delta}\right).$$

We conclude our discussion of this special case with $\|\mathcal{A}\| = 2$ by exploiting the result in Proposition 4 to derive upper and lower bounds for the value function.

Proposition 6. The value function $\tilde{\mathcal{G}}_{ij}$ satisfies

$$\tilde{\mathcal{G}}_{ij}(\delta) \leq \tilde{\mathcal{G}}_{ij}(\delta) \leq \tilde{\mathcal{G}}_{ij}(0)(1-\delta) + \tilde{\mathcal{G}}_{ij}(1)\delta$$

for all $\delta \in (0, 1)$.

Furthermore,

$$\max_{\delta \in (0, 1)} \{\tilde{\mathcal{G}}_{ij}(\delta) - \tilde{\mathcal{G}}_{ij}(\delta)\} = \tilde{\mathcal{G}}_{ij}(\hat{\delta}_{ij}) - \tilde{\mathcal{G}}_{ij}(\hat{\delta}_{ij}).$$

4.3.3. General Case: $|\mathcal{A}| \geq 2$. Let us now turn to the general case in which the set \mathcal{A} includes an arbitrary but finite number of actions. Our derivation of the value function $\tilde{\mathcal{G}}(\delta)$ in (12) is obtained based on the solution derived in the previous section. For each pair of actions $\{a_i, a_j\} \in \mathcal{A}$, the function $\tilde{\mathcal{G}}_{ij}(\delta)$ in Proposition 5 is the value function of a problem in which only actions a_i and a_j are available. It follows that $\tilde{\mathcal{G}}(\delta) \geq \tilde{\mathcal{G}}_{ij}(\delta)$ for all $\delta \in [0, 1]$ and so

$$\tilde{\mathcal{G}}(\delta) \geq \tilde{V}(\delta) := \max_{\{a_i, a_j\} \in \mathcal{A}} \{\tilde{\mathcal{G}}_{ij}(\delta)\},$$

where $\tilde{V}(\delta)$ is the pointwise maximum of the functions $\tilde{\mathcal{G}}_{ij}(\delta)$. Our main result in this section establishes that the inequality is, in fact, an equality; that is, $\tilde{\mathcal{G}}(\delta) = \tilde{V}(\delta)$. To prove this, we show that the function $\tilde{V}(\delta)$ satisfies the QVI conditions so that we can invoke the verification Theorem 1. To this end, we first show that $\tilde{V}(\delta)$ satisfies all three QVI conditions in (15).

Proposition 7. For all $\delta \in [0, 1]$, we have that $\tilde{V}(\delta) \geq \tilde{\mathcal{G}}(\delta)$. Also, there exists a finite set $N_{\tilde{V}} \subseteq [0, 1]$ such that $\mathcal{H}\tilde{V}(\delta) \leq 0$ and $(\tilde{V}(\delta) - \tilde{\mathcal{G}}(\delta))\mathcal{H}\tilde{V}(\delta) = 0$ for all $\delta \in [0, 1] \setminus N_{\tilde{V}}$.

The attentive reader might have noticed that the result in Proposition 7 is not enough to invoke the verification Theorem 1. The reason is that, besides verifying the QVI conditions, we also need to show that the function $\tilde{V}(\delta)$ is sufficiently smooth and belongs to the set $\tilde{\mathcal{C}}^2$ (see Equation (13)). We formalize this condition in the following result.

Theorem 2. The function $\tilde{V}(\delta) = \max_{\{a_i, a_j\} \in \mathcal{A}} \{\tilde{\mathcal{G}}_{ij}(\delta)\}$ is in $\tilde{\mathcal{C}}^2$. As a result, $\tilde{\mathcal{G}}(\delta) = \tilde{V}(\delta)$.

It is worth noticing that the previous theorem shows that the complexity of the diffusion optimal stopping problem grows only quadratically with the cardinality of the action set \mathcal{A} . In fact, Theorem 2 reveals that solving a problem with $|\mathcal{A}|$ actions is equivalent to solving a collection of $|\mathcal{A}|(|\mathcal{A}| - 1)$ problems each with only two actions.

To illustrate the result in Theorem 2 and the resulting optimal policy, let us consider the example in Figure 4 in which the set \mathcal{A} has four actions. The left panel depicts all six functions $\{\tilde{\mathcal{G}}_{ij}(\delta) : \{a_i, a_j\} \in \mathcal{A}\}$, and the right panel depicts the function $\tilde{V}(\delta) := \max_{\{a_i, a_j\} \in \mathcal{A}} \{\tilde{\mathcal{G}}_{ij}(\delta)\}$.

After a quick inspection, we can check that, in this example, there exist two (nonunique) thresholds $\underline{\delta}$ and $\bar{\delta}$ such that

$$\tilde{V}(\delta) = \begin{cases} \tilde{\mathcal{G}}_{12}(\delta) & \text{if } 0 \leq \delta \leq \underline{\delta} \\ \tilde{\mathcal{G}}_{23}(\delta) & \text{if } \underline{\delta} \leq \delta \leq \bar{\delta} \\ \tilde{\mathcal{G}}_{34}(\delta) & \text{if } \bar{\delta} \leq \delta \leq 1. \end{cases}$$

Furthermore, at $\delta = \underline{\delta}$, the functions $\tilde{\mathcal{G}}_{12}(\delta)$ and $\tilde{\mathcal{G}}_{23}(\delta)$ meet smoothly because $\tilde{\mathcal{G}}_{12}(\underline{\delta}) = \tilde{\mathcal{G}}_{23}(\underline{\delta}) = \tilde{\mathcal{R}}_2(\underline{\delta})$. A similar smooth pasting occurs at $\delta = \bar{\delta}$ because $\tilde{\mathcal{G}}_{23}(\bar{\delta}) = \tilde{\mathcal{G}}_{34}(\bar{\delta}) = \tilde{\mathcal{R}}_3(\bar{\delta})$. As a result, because each of the functions $\tilde{\mathcal{G}}_{ij}(\delta)$ is in $\tilde{\mathcal{C}}^2$ by Proposition 5, it follows that $\tilde{V}(\delta)$ is also in $\tilde{\mathcal{C}}^2$. Note also that an optimal policy is given by a sequence of thresholds that define the continuation and intervention regions. In this example, we have that

Continuation: $\mathcal{C}^A = (\underline{\delta}_{12}, \bar{\delta}_{12}) \cup (\underline{\delta}_{23}, \bar{\delta}_{23}) \cup (\underline{\delta}_{34}, \bar{\delta}_{34})$.

Intervention: $\mathcal{I}^A = [0, \underline{\delta}_{12}] \cup [\bar{\delta}_{12}, \underline{\delta}_{23}] \cup [\bar{\delta}_{23}, \underline{\delta}_{34}] \cup [\bar{\delta}_{34}, 1]$,

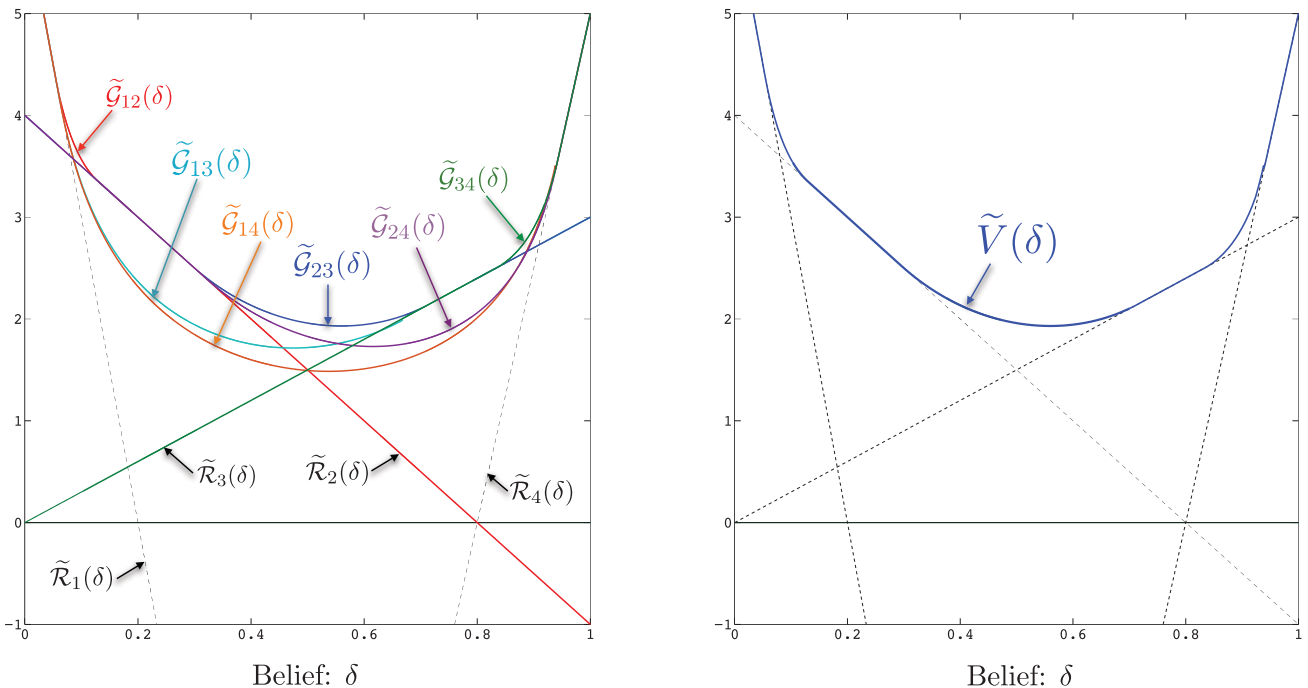
where the thresholds $\underline{\delta}_{ij}$ and $\bar{\delta}_{ij}$ are defined in Proposition 5.

5. Nonasymptotic Experimentation Policies

In this section, we discuss how to interpret the asymptotic analysis developed in the previous section to construct experimentation policies that can be used in an arbitrary instance. Recall from Definition 1 that a policy consists of two components: (a) an intervention region \mathcal{I} that defines the set of beliefs δ at which the decision maker stops the experimentation process and selects an optimal action $a^* \in \mathcal{A}^*(\delta)$, and (b) an experimentation policy $\pi(\delta) \in \mathcal{E}$ that identifies the experiment that the DM should conduct at each δ in the continuation region $\mathcal{C} = \mathcal{I}^c$.

The asymptotic analysis of the previous section produces an experimentation strategy $\pi(\delta) = \tilde{\mathcal{E}}^A$ and an intervention region \mathcal{I}^A defined in Corollary 1 and Proposition 5, respectively. However, we cannot implement these policies directly because they are computed in terms of the nonprimitive quantities $\mathcal{Q}(x, \mathcal{E})$, $\alpha(x, \mathcal{E}, \theta_0)$ and $\alpha(x, \mathcal{E}, \theta_1)$ appearing in Assumption 1. Therefore, in order to recover a solution to an arbitrary instance of the problem from the asymptotic analysis of the previous section, we need to derive the values of

Figure 4. (Color online) Example in Which the Offer Set \mathcal{O} Includes Four Products



Notes. The left panel depicts the value functions $\tilde{\mathcal{G}}_{ij}(\delta)$ derived in Proposition 5. The right panel depicts the function $V(\delta) := \max_{\{i,j\} \in \mathcal{O}} \{\tilde{\mathcal{G}}_{ij}(\delta)\}$. Data: $\mathcal{R}_1(\delta) = 6 - 30\delta$, $\mathcal{R}_2(\delta) = 4 - 5\delta$, $\mathcal{R}_3(\delta) = 3\delta$, $\mathcal{R}_4(\delta) = -20 + 25\delta$, $r = 1$, and $\bar{\sigma} = 2$.

Downloaded from informs.org by [212.36.194.19] on 15 April 2024, at 21:31. For personal use only, all rights reserved.

$Q(x, \mathcal{E})$, $\alpha(x, \mathcal{E}, \theta_0)$ and $\alpha(x, \mathcal{E}, \theta_1)$ from the primitives of the model, namely, from the values of Λ , $Q(x, \mathcal{E}, \theta_0)$, and $Q(x, \mathcal{E}, \theta_1)$.

In some settings, this derivation can be done directly by imposing a specific parametric structure in the definitions of $Q(x, \mathcal{E}, \theta_0)$ and $Q(x, \mathcal{E}, \theta_1)$. The idea in these settings is that the parametric structure is used to capture a distinctive feature of the problem at hand, which, in turn, would determine the asymptotic regime of interest. Let us illustrate this point with a concrete example.

Example 3. Consider a setting in which the primitives $Q(x, \mathcal{E}, \theta)$ are known and assumed to be continuously differentiable in θ for θ in some open neighborhood that contains the two hypotheses $\Theta = \theta_0$ and $\Theta = \theta_1$. Suppose that we are interested in a setup in which a distinctive characteristic of the problem is that the two hypotheses are hard to distinguish. We can model this feature by setting $\theta_1 = \theta_0 + \xi/\sqrt{k}$ for some fixed scalar ξ . Using a first order Taylor expansion, it follows that $Q(x, \mathcal{E}, \theta_1) = Q(x, \mathcal{E}, \theta_0) + Q_\theta(x, \mathcal{E}, \theta_0) \xi/\sqrt{k} + o(k^{1/2})$, where $Q_\theta(x, \mathcal{E}, \theta)$ is the partial derivative of $Q(x, \mathcal{E}, \theta)$ with respect to θ . Under this specific parameterization of the problem, we can now apply the asymptotic analysis of the previous section (by letting k go to infinity) to derive the corresponding values of $Q(x, \mathcal{E})$, $\alpha(x, \mathcal{E}, \theta_0)$ and $\alpha(x, \mathcal{E}, \theta_1)$. In this case, it is not hard to see that

$$Q(x, \mathcal{E}) = Q(x, \mathcal{E}, \theta_0), \quad \alpha(x, \mathcal{E}, \theta_0) = 0 \quad \text{and} \\ \alpha(x, \mathcal{E}, \theta_1) = Q_\theta(x, \mathcal{E}, \theta_0) \xi. \quad \diamond$$

In Section 6, we consider at length a concrete application related to crowd voting in which $Q(x, \mathcal{E}, \theta)$ are viewed as choice probabilities governed by an MNL model. In this context, similarly to Example 3, we also derive the quantities $Q(x, \mathcal{E})$ and $\alpha(x, \mathcal{E}, \theta)$ not only for the case of indistinguishable hypotheses, but also for the case in which the experiment outcomes are very noisy.

The previous example provides some insights into how one can leverage some concrete knowledge about the structure of the problem to identify the proper asymptotic regime to use. However, this approach does not generalize in an obvious way to an arbitrary setting for which such knowledge is not available. In what follows, we propose a methodology that does not rely on any additional information beyond the values of Λ , $Q(x, \mathcal{E}, \theta_0)$, and $Q(x, \mathcal{E}, \theta_1)$.

Combining the asymptotic scalings in Equations (8) and (9), we have that the input parameters Λ^k , $Q^k(x, \mathcal{E}, \theta_0)$ and $Q^k(x, \mathcal{E}, \theta_1)$ satisfy the following relationship for k large

$$\sqrt{\Lambda^k} \left(\frac{Q^k(x, \mathcal{E}, \theta)}{Q(x, \mathcal{E})} - 1 \right) \approx \alpha(x, \mathcal{E}, \theta). \quad (19)$$

Furthermore, going back to the conditions that define the asymptotic regime in Corollary 1, we see that the value of $Q(x, \mathcal{E})$ is such that the quantity $\frac{Q^k(x, \mathcal{E}, \theta)}{Q(x, \mathcal{E})} - 1$ converges to zero at a rate of $1/\sqrt{k}$ uniformly in $\mathcal{E} \in \mathcal{E}$, $x \in \mathcal{X}$ and $\theta = \theta_0, \theta_1$. Thus, in a nonasymptotic regime, we can reinterpret this condition as one that requires $Q(x, \mathcal{E})$ to be as close to $Q(x, \mathcal{E}, \theta)$ as possible for all $\mathcal{E} \in \mathcal{E}$, $x \in \mathcal{X}$ and $\theta = \theta_0, \theta_1$. In other words, we can represent this problem as an optimization problem that minimizes the “distance” between $Q(x, \mathcal{E})$ and $Q(x, \mathcal{E}, \theta)$. We propose the following min-max formulation to compute $Q(x, \mathcal{E})$:

$$\text{For every } \mathcal{E} \in \mathcal{E} \text{ solve: } \min_{Q \geq 0} \max_{\theta \in \{\theta_0, \theta_1\}} \max_{x \in \mathcal{X}} \left| \frac{Q(x, \mathcal{E}, \theta)}{Q(x, \mathcal{E})} - 1 \right| \\ \text{subject to } \sum_{x \in \mathcal{X}} Q(x, \mathcal{E}) = 1. \quad (20)$$

After computing the value of $Q(x, \mathcal{E})$, we can obtain the values of $\alpha(x, \mathcal{E}, \theta_0)$ and $\alpha(x, \mathcal{E}, \theta_1)$ using (19), that is,

$$\alpha(x, \mathcal{E}, \theta) \approx \sqrt{\Lambda} \left(\frac{Q(x, \mathcal{E}, \theta)}{Q(x, \mathcal{E})} - 1 \right), \quad \text{for } \theta = \theta_0, \theta_1. \quad (21)$$

The following proposition establishes the consistency between the value of the probability kernel $Q(x, \mathcal{E})$ computed in (20) and the corresponding asymptotic limit.

Proposition 8. Consider a sequence of probability distributions $\{Q^k(x, \mathcal{E}, \theta); \mathcal{E} \in \mathcal{E}, x \in \mathcal{X}, \theta = \theta_0, \theta_1\}$ satisfying the condition in Assumption 1. In particular, $Q^k(x, \mathcal{E}, \theta) \rightarrow Q(x, \mathcal{E})$ for some probability kernel $Q(x, \mathcal{E})$ for $\theta = \theta_0, \theta_1$. Moreover, for each k and $Q^k(x, \mathcal{E}, \theta)$, let $Q^k(x, \mathcal{E})$ be the corresponding solution to (20). Then, $Q^k(x, \mathcal{E})$ converges to $Q(x, \mathcal{E})$ as $k \uparrow \infty$ for all $\mathcal{E} \in \mathcal{E}$ and $x \in \mathcal{X}$.

Let us now turn to the issue of how to adapt the asymptotic solutions to derive implementable policies. Equations (20) and (21) allow us to compute the values of $Q(x, \mathcal{E})$, $\alpha(x, \mathcal{E}, \theta_0)$ and $\alpha(x, \mathcal{E}, \theta_1)$ that are needed to derive the asymptotic strategy $\pi(\delta) = \tilde{\mathcal{E}}^A$ and \mathcal{I}^A . From Corollary 1, we have that

$$\tilde{\mathcal{E}}^A = \arg \max_{\mathcal{E} \in \mathcal{E}} \left\{ \sum_{x \in \mathcal{X}_{\mathcal{E}}} (\alpha(x, \mathcal{E}, \theta_1) - \alpha(x, \mathcal{E}, \theta_0))^2 Q(x, \mathcal{E}) \right\}.$$

On the other hand, the value of \mathcal{I}^A is obtained from our diffusion analysis of the optimal stopping problem combining the results in Proposition 5 and Theorem 2. The volatility $\tilde{\sigma}$ of the underlying diffusion process is the one identified in Proposition 2, that is,

$$\tilde{\sigma}^2 = \sum_{x \in \mathcal{X}_{\tilde{\mathcal{E}}^A}} (\alpha(x, \tilde{\mathcal{E}}^A, \theta_1) - \alpha(x, \tilde{\mathcal{E}}^A, \theta_0))^2 Q(x, \mathcal{E}).$$

We use this asymptotic solution $(\tilde{\mathcal{E}}^A, \mathcal{I}^A)$ to propose two concrete approximation policies.

- **Asymptotic Policy (A)**: This policy implements directly the strategy $(\tilde{\mathcal{E}}^A, \mathcal{I}^A)$.

- **Maximum Volatility Policy (MV)**: This policy uses the same intervention region \mathcal{I}^A as the asymptotic policy. On the other hand, in terms of experimentation, the MV policy reinterprets the solution in Corollary 1 and, for each δ in the continuation region, selects the experiment that maximizes the instantaneous volatility, that is, $\pi^{\text{MV}}(\delta) = \mathcal{E}^{\text{MV}}(\delta)$, where

$$\mathcal{E}^{\text{MV}}(\delta) = \arg \max_{\mathcal{E} \in \mathcal{E}} \left\{ \mathbb{E}_0 \left[\frac{(1 - \mathcal{L}(\mathcal{E}))^2}{\delta + (1 - \delta)\mathcal{L}(\mathcal{E})} \right] \right\}. \quad (22)$$

(The subscript “MV” is mnemonic for maximum volatility.)

Note that the asymptotic policy suggests a static experimentation, and maximum volatility offers a dynamic experimentation function of the current belief. However, it should be clear from our previous discussion that both of these policies are asymptotically equal and optimal in the limiting regime defined by Equations (8) and (9). It is also worth noticing that, in contrast to the derivation of an optimal experimentation policy in Equation (6) that requires full knowledge of the value function, the MV policy can be computed directly using only the knowledge of the likelihood function $\mathcal{L}(x, \mathcal{E})$. This, of course, simplifies significantly its computational complexity.

In Section 7, we conduct a set of numerical experiments to test the performance of our proposed policies using a concrete application in the context of new product introduction that we present in the next section. We conclude this section with a remark on how to extend some of the insights that have developed to the problem of designing the type of experiments that the DM can use.

5.1. A Remark on the Optimal Design of Experiments

In some applications (such as the assortment selection problem discussed in the next section), the decision maker has some degree of control over the design of the set \mathcal{E} of available experiments. In such cases, observe that the optimization in (22) that defines $\mathcal{E}^{\text{MV}}(\delta)$ can be reformulated over a more abstract set \mathbb{L} of likelihood ratios, in which each $\mathcal{L} \in \mathbb{L}$ corresponds to an experiment \mathcal{E} . As a result, the optimization problem that defines the maximum volatility policy is given by

$$\max_{\mathcal{L} \in \mathbb{L}} \mathbb{E}_\delta \left[\left(\frac{1 - \mathcal{L}}{\delta + (1 - \delta)\mathcal{L}} \right)^2 \right], \quad \text{or equivalently,}$$

$$\max_{\mathcal{L} \in \mathbb{L}} \mathbb{E}_0 \left[\frac{(1 - \mathcal{L})^2}{\delta + (1 - \delta)\mathcal{L}} \right],$$

where $\mathbb{E}_0[\cdot]$ denotes the expectation under the probability measure $\mathbb{P}_0(\cdot)$.

Depending on the nature of the set \mathbb{L} , this optimization problem can be cast as a Tchebycheff moment problem. Consider the following setting in which the DM can design experiments that correspond to any possible likelihood ratio \mathcal{L} as long as \mathcal{L} is bounded by two given quantities $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$. In this case, the following result holds.

Proposition 9. Suppose that $\mathbb{L} = \{\mathcal{L} : \mathbb{E}_0[\mathcal{L}] = 1 \text{ and } \underline{\mathcal{L}} \leq \mathcal{L} \leq \bar{\mathcal{L}}\}$ for two nonnegative scalars $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$ and let

$$\mathcal{L}^* = \arg \max_{\mathcal{L} \in \mathbb{L}} \mathbb{E}_0 \left[\frac{(1 - \mathcal{L})^2}{\delta + (1 - \delta)\mathcal{L}} \right].$$

Then, \mathcal{L}^* is a random variable with a two-point distribution with mass at $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$.

Proof. The result follows from noticing that the function $(1 - \ell)^2 / (\delta + (1 - \delta)\ell)$ is convex, and so an optimal solution is a two-point distribution with mass at $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$. \square

The solution in Proposition 9 suggests that the decision maker should select an experimentation policy that maximizes the range of the likelihood function. In Section 7, we explore this idea and propose a variation of the maximum volatility policy that incorporates this “maximum range” condition and shows very good numerical performance.

6. Illustrative Example: New Product Introduction

We discuss in this section a concrete application of the methodology and results presented in the previous sections in the context of a new product introduction problem. The literature on the topic is quite broad (see the recent work of Sunar et al. 2019 and references therein). In particular, we consider an environment in which the experimentation outcomes are the result of a consumer voting process driven by an MNL. Our objective in developing this example is twofold. First, we use it to provide some specific details on how to formulate and derive our proposed asymptotic approximation policies discussed in the previous section. As a by-product of this discussion, we also show how to obtain diffusion approximations for a belief process that is governed by an MNL model using two different types of asymptotic regimes. Our second objective is to use this concrete example in Section 7 to conduct a set of numerical experiments to test the quality of our proposed methodology.

6.1. Model Setup

The specific setting that we consider is as follows. Consider a seller (or firm) that is contemplating the

possibility of introducing a new product (or products) into the marketplace. In the process of developing these new products, the seller has prototyped n different versions and would like to decide which is the right subset to commercialize if any. These prototypes differ in terms of some specific set of attributes which might include their price and quality as well as launching and manufacturing costs to name a few. We assume that the intrinsic utility that a consumer assigns to version $i \in [n]$ is equal to $u_i(\Theta)$, where $\Theta > 0$ is some unknown real parameter.

Example 4 (Linear Utilities). A popular modeling approach is to assume that the utilities $u_i(\Theta)$ are linear in the unknown parameter Θ . For instance, we can have $u_i(\Theta) = q_i - p_i\Theta$, where q_i and p_i are product i 's quality and price, respectively. In this case, Θ measures consumers' price sensitivity. \diamond

The seller is uncertain about market conditions and does not know the value of the parameter Θ . In an attempt to reduce the risk of launching the wrong version(s), the seller sets up an online voting system in which potential customers (those visiting the seller's website) can vote for the different prototypes. For simplicity, we assume that each voter votes for at most one version, and the seller only tracks the cumulative number of votes for each one. This voting phase occurs before the seller decides to launch a product and has the potential of offering a win-win situation for both the consumer and the seller. As we show later, it is not necessarily optimal for the seller to display the entire set $[n]$ during the voting phase. Hence, we assume that the seller selects a subset \mathcal{E} of prototypes to show each voter during the voting phase. We call \mathcal{E} the *display set* and let $|\mathcal{E}|$ be its cardinality. To keep some consistency between the notation in this and the previous sections, we note that, in the most general case, both the set of experiments \mathcal{E} and available actions \mathcal{A} coincide with the power set of $[n]$, that is, $\mathcal{E} = \mathcal{A} = 2^{[n]}$. In some cases, however, one might need to restrict the set of experiments and actions. For instance, if the number of prototypes is large, then it might be impractical to display the entire menu, and experimentation should be restricted to display sets of a given cardinality. Similarly, it is also possible that the seller is constrained in the number of versions that it can launch.

Voters arrive according to a Poisson process with rate Λ and vote for one alternative from the display set according to a multinomial choice model. Specifically, a voter who observes a display set \mathcal{E} assigns to each version $i \in \mathcal{E}$ a utility $U_i(\Theta) = u_i(\Theta) + \varepsilon_i$, where $\{\varepsilon_i : i \in \mathcal{E}\}$ are idiosyncratic utility shocks that are independent and identically distributed according to a Gumbel distribution with mean zero and variance

$\text{Var}[\varepsilon] = \pi^2/(6\mu^2)$ for some fixed constant $\mu > 0$. It follows that a utility-maximizing voter votes for version $i \in \mathcal{E}$ with probability

$$Q(i, \mathcal{E}, \Theta) := \mathbb{P}(U_i(\Theta) \geq U_j(\Theta), \forall j \in \mathcal{E}) = \frac{\exp(\mu u_i(\Theta))}{\sum_{j \in \mathcal{E}} \exp(\mu u_j(\Theta))}. \quad (23)$$

Note that our formulation allows for the possibility that a voter might end up not selecting any of the available options. To model this no-vote option, we simply include version zero with quality, price, and intrinsic utility equal to zero, $u_0(\Theta) = 0$. In what follows, we assume that every display set \mathcal{E} includes the nonpurchase option.

We assume that the seller has a prior belief about the value of Θ that can take one of two possible values $\{\theta_0, \theta_1\}$, and its prior is that $\Theta = \theta_0$ with probability $\delta \in (0, 1)$. We let $u_i(\theta_0)$ and $u_i(\theta_1)$ denote voters' intrinsic utilities under these two hypotheses for $i \in [n]$ and define the likelihood ratio function by

$$\mathcal{L}(i, \mathcal{E}) := \frac{Q(i, \mathcal{E}, \theta_1)}{Q(i, \mathcal{E}, \theta_0)} \quad \forall i \in \mathcal{E}. \quad (24)$$

We complete the description of the model by specifying the seller's objective function. As in the general case, we assume that there exists a piecewise linear function $G(\delta)$ (see Equation (1)) that represents the seller's expected payoff as function of its belief δ . The seller's optimization problem is given by

$$\begin{aligned} \Pi(\delta) = \sup_{(\pi, \mathcal{I})} \mathbb{E}_\delta[e^{-r\delta_\tau} G(\delta_\tau)], \quad \text{subject to} \\ \tau = \inf\{t > 0 : \delta_t \in \mathcal{I}\}. \end{aligned} \quad (25)$$

Recall that a policy is defined by an experimentation policy π that determines the collection of display sets $\{\mathcal{E}_t \in \mathcal{E} : 0 \leq t \leq \tau\}$ to use throughout the voting process and an intervention region \mathcal{I} that defines the duration of the voting campaign.

Remark 2 (Payoffs from Sales). To illustrate a concrete example of the piecewise linear payoff function $G(\delta)$ in the context of new product introduction, consider the case in which the seller is interested in maximizing the expected discounted value of the cash flows generated by the sales that occur after time τ . Specifically, at time τ , the seller stops the voting process and selects a subset $\mathcal{A} \in \mathcal{A}$ of products to launch based on the available information at this time. Suppose consumers arrive according to a Poisson process of rate Λ_s and make buying decision according

to the same MNL model that governs the voting process. Under this assumption, the seller's expected discounted payoff is given by

$$\begin{aligned} \mathcal{R}(\delta_\tau, \mathcal{A}) &:= \mathbb{E} \left[\sum_{i \in \mathcal{A}} \int_\tau^\infty e^{-r(t-\tau)} (p_i - c_i) dS_{it} - K_i \mid \mathcal{F}_\tau \right] \\ &= \sum_{i \in \mathcal{A}} \left[\frac{(p_i - c_i)}{r} \Lambda_s \mathbb{E}[Q(i, \mathcal{A}, \Theta) \mid \mathcal{F}_\tau] - K_i \right] \\ &= \phi(\mathcal{A}) + \beta(\mathcal{A}) \delta_\tau, \end{aligned}$$

where p_i , c_i , and K_i are the per-unit price, manufacturing cost, and fixed launching cost of product $i \in \mathcal{S}$, respectively, and

$$\begin{aligned} \phi(\mathcal{A}) &:= \sum_{i \in \mathcal{A}} \left[\frac{(p_i - c_i)}{r} \Lambda_s Q(i, \mathcal{A}, \theta_1) - K_i \right] \quad \text{and} \\ \beta(\mathcal{A}) &:= \sum_{i \in \mathcal{A}} \left[\frac{(p_i - c_i)}{r} \Lambda_s (Q(i, \mathcal{A}, \theta_0) - Q(i, \mathcal{A}, \theta_1)) \right]. \end{aligned}$$

In the case that all products are discarded, one can assume the seller receives a fixed payoff \mathcal{R}_0 (possibly zero), which captures the opportunity cost of the business. Finally, the seller's payoff function in this case is given by $G(\delta) = \max \{ \mathcal{R}(\delta, \mathcal{A}) : \mathcal{A} \in \mathcal{A} \}$. \diamond

6.2. Asymptotic Approximation

We move now to apply the results in Section 4 to approximate the optimization in (25) by a diffusion control problem. In order to invoke the weak convergence result in Proposition 2, we need to specify an asymptotic regime under which the MNL choice probabilities satisfy the condition in Equation (8). In what follows, we propose two concrete alternatives, each capturing a different type of unformativeness associated with the voting process.

6.2.1. Noisy Preferences. Motivated by the issue of low-quality data that has been reported in the context of online learning applications and advertising (Lewis and Rao 2015, Kohavi and Thomke 2017), we consider a regime in which the variance of the MNL idiosyncratic shocks in the k^{th} instance of the problem grows proportionally with k , namely, $\text{Var}[\varepsilon^k] = k \pi^2 / (6 \mu^2)$. In other words, this asymptotic regime is one in which votes—and the information they contain—become more and more noisy as k grows large.

Under this scaling, one can show that the choice probability $Q^k(i, \mathcal{E}, \theta)$ in (23) can be written as

$$Q^k(i, \mathcal{E}, \theta) = \frac{1}{|\mathcal{E}|} \left[1 + \frac{\mu}{|\mathcal{E}| \sqrt{k}} \sum_{j \in \mathcal{E}} (u_i(\theta) - u_j(\theta)) + o(k^{-1/2}) \right], \quad (26)$$

which satisfies the requirements in Assumption 1 with

$$Q(i, \mathcal{E}) = \frac{1}{|\mathcal{E}|} \quad \text{and} \quad \alpha(i, \mathcal{E}, \theta) = \mu (u_i(\theta) - \bar{u}(\mathcal{E}, \theta))$$

$$\text{where } \bar{u}(\mathcal{E}, \theta) := \frac{1}{|\mathcal{E}|} \sum_{j \in \mathcal{E}} u_j(\theta).$$

Recall that, under our asymptotic scaling, voters arrive according to a Poisson process N_t^k with intensity $\Lambda^k = k \Lambda$ in the k^{th} instance of the problem. Given this scaling of $\text{Var}[\varepsilon^k]$ and Λ^k , we can use the result in Proposition 2 to obtain the following corollary.

Corollary 3. Let $\Delta u_i := u_i(\theta_1) - u_i(\theta_0)$ and $\Delta \bar{u}(\mathcal{E}) := \bar{u}(\mathcal{E}, \theta_1) - \bar{u}(\mathcal{E}, \theta_0)$. Suppose the seller uses a static display policy $\mathcal{E}_t = \mathcal{E}$ during the voting process. Then, the belief process δ_t^k converges weakly to the solution of the SDE

$$\begin{aligned} d\tilde{\delta}_t &= \tilde{\sigma}(\mathcal{E}) \tilde{\delta}_t (1 - \tilde{\delta}_t) dW_t, \quad \text{where} \\ \tilde{\sigma}^2(\mathcal{E}) &= \frac{\Lambda \mu^2}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} (\Delta u_i - \Delta \bar{u}(\mathcal{E}))^2, \end{aligned}$$

and W_t is a Wiener process.

Combining this result together with the maximum volatility principle in Corollary 1, we can now identify an optimal display set in this asymptotic regime under consideration, namely,

$$\tilde{\mathcal{E}}_{\text{NP}}^A = \arg \max_{\mathcal{E} \in \mathcal{E}} \left\{ \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} (\Delta u_i - \Delta \bar{u}(\mathcal{E}))^2 \right\}. \quad (27)$$

(The subscript “NP” stands for noisy preferences regime.)

Without loss of generality, let us index the prototypes in ascending order of Δu so that $\Delta u_1 \leq \Delta u_2 \leq \dots \leq \Delta u_n$. Also, for $0 \leq i \leq j \leq n$, let us define the display set

$$\mathcal{E}[i, j] := \{0\} \cup \{1, \dots, i\} \cup \{j, \dots, n\}, \quad (28)$$

which includes the nonpurchase option together with the first i prototypes with the lowest values of Δu and the $n - j + 1$ prototypes with the highest values of Δu .

Proposition 10. Let $\tilde{\mathcal{E}}_{\text{NP}}^A$ be a solution to (27). Then, there exist integers n_1 and n_2 with $0 \leq n_1 < n_2 \leq n$ such that $\tilde{\mathcal{E}}_{\text{NP}}^A = \mathcal{E}[n_1, n_2]$. Furthermore, in the special case that all the $\{\Delta u_i\}$ have the same sign (i.e., $\Delta u_i \geq 0$ or $\Delta u_i \leq 0$), then $\tilde{\mathcal{E}}_{\text{NP}}^A$ consists of a single prototype $i^* = \arg \max \{ |\Delta u_i| : i \in \mathcal{S} \}$ together with the nonpurchase option zero, that is, $\tilde{\mathcal{E}}_{\text{NP}}^A = \{0, i^*\}$.

An important corollary of Proposition 10 is that instead of solving (27) over the power set of \mathcal{E} , we can restrict ourselves to the much simpler problem of maximizing the volatility of the belief process over the significantly smaller class of display sets $\{\mathcal{E}[i, j] : 0 \leq i \leq j \leq n\}$, which has a cardinality of $O(n^2)$.

Example 5 (Example 4 Revisited). Suppose the intrinsic utility of product i is equal to $u_i(\Theta) = q_i - p_i \Theta$; then, $\Delta u_i = p_i(\theta_0 - \theta_1)$. If all the $\{p_i\}$ are of the same sign, for example, if they correspond to the prices of the products, then the $\{\Delta u_i\}$ are also of the same sign, and the optimal display set $\tilde{\mathcal{E}}_{\text{NP}}^A$ includes a single prototype, namely, the one with the highest price. \diamond

6.2.2. Asymptotically Indistinguishable Hypotheses.

An alternative regime in which we can apply the asymptotic analysis of Section 4 corresponds to the case in which the values of θ_0 and θ_1 become indistinguishable as k grows large. To be precise, let us consider the case in which $u_i(\theta_1) = u_i(\theta_0) + \xi_i/\sqrt{k}$ for $i \in [n]$, where $\{\xi_1, \xi_2, \dots, \xi_n\}$ are fixed constants independent of k . Under this scaling, the choice probability $Q^k(i, \mathcal{E}, \theta)$ in (23) admits the following representation:

$$Q^k(i, \mathcal{E}, \theta_0) = \frac{v_i}{\sum_{j \in \mathcal{E}} v_j} \quad \text{and}$$

$$Q^k(i, \mathcal{E}, \theta_1) = \frac{v_i}{\sum_{j \in \mathcal{E}} v_j} \left[1 + \frac{1}{\sqrt{k}} \frac{\sum_{j \in \mathcal{E}} v_j (\xi_i - \xi_j)}{\sum_{j \in \mathcal{E}} v_j} + o(k^{-1/2}) \right], \quad (29)$$

where $v_i := \exp(\mu u_i(\theta_0))$. It follows that these choice probabilities satisfy the conditions in Assumption 1 with

$$Q(i, \mathcal{E}) = \frac{v_i}{\sum_{j \in \mathcal{E}} v_j}, \quad \alpha(i, \mathcal{E}, \theta_0) = 0 \quad \text{and}$$

$$\alpha(i, \mathcal{E}, \theta_1) = \sum_{j \in \mathcal{E}} (\xi_i - \xi_j) Q(j, \mathcal{E}).$$

From Corollary 1, the optimal display set in this asymptotic regime is given by

$$\tilde{\mathcal{E}}_{\text{IH}}^A = \arg \max_{\mathcal{E} \in \mathcal{E}} \left\{ \sum_{i \in \mathcal{E}} (\alpha(i, \mathcal{E}, \theta_1))^2 Q(i, \mathcal{E}) \right\}. \quad (30)$$

(The subscript “IH” stands for indistinguishable hypotheses regime.)

To get some intuition about $\tilde{\mathcal{E}}_{\text{IH}}^A$, consider an arbitrary display set $\mathcal{E} \in \mathcal{E}$ and let $\xi(\mathcal{E})$ be random variables taking values in $\{\xi_1, \xi_2, \dots, \xi_n\}$ with probability distribution $Q(i, \mathcal{E})$. (In this definition, we assume that $Q(i, \mathcal{E}) = 0$ if $i \notin \mathcal{E}$.) Then, (30) can be rewritten as

$$\tilde{\mathcal{E}}_{\text{IH}}^A = \arg \max_{\mathcal{E} \in \mathcal{E}} \{\text{Var}[\xi(\mathcal{E})]\}.$$

Remark 3. It is worth noticing that the asymptotic regime in which the two alternative hypotheses $\Theta = \theta_0$ and $\Theta = \theta_1$ are asymptotically indistinguishable does not imply that the DM optimization problem becomes trivial in the limit. To see this, let us consider the payoff structure discussed in Remark 2, in which $\mathcal{R}(\delta, \mathcal{A})$ is the discounted payoff that the DM expects to collect if the DM launches assortment \mathcal{A} when the DM’s

belief is δ . Under the scaling in (29), it is not hard to show that, for the k^{th} instance,

$$\mathcal{R}^k(1, \mathcal{A}) - \mathcal{R}^k(0, \mathcal{A}) = \sum_{i \in \mathcal{A}} \left[\frac{(p_i - c_i)}{r} \frac{\Lambda_s^k}{\sqrt{k}} \alpha(i, \mathcal{A}, \theta_1) Q(i, \mathcal{A}) \right] + o(k^{-1/2}),$$

where Λ_s^k is the selling rate after launching. Thus, depending on the rate of growth of Λ_s^k in k , there is a nonnegligible difference in payoffs between the two hypotheses, and so it is in the DM’s best interest to try to learn which one holds true. \diamond

7. Numerical Experiments

In this section, we conduct a set of numerical experiments to assess the quality of our methodology using the application discussed in the previous section. In particular, we are interested in investigating the performance of our proposed asymptotic and maximum volatility policies introduced in Section 5. Online Appendix C contains some additional computational experiments.

7.1. Optimality Gap

In our first set of computational experiments, we numerically evaluate the optimality gap of the asymptotic and maximum volatility policies with respect to an optimal policy using the *noisy preferences* model in Section 6.2.1. We let $\Pi^A(\delta)$, $\Pi^{\text{MV}}(\delta)$ and $\Pi(\delta)$ denote the value functions generated by the A, MV, and optimal policies, respectively, and define the optimality gap of these policies by

$$\Delta \Pi^j := \max_{\delta \in (0,1)} \left\{ \frac{\Pi(\delta) - \Pi^j(\delta)}{\Pi(\delta)} \right\}, \quad j = A, \text{MV}.$$

We measure $\Delta \Pi^A$ and $\Delta \Pi^{\text{MV}}$ using a set of 500 random instances of the problem. Specifically, we consider a problem with $n = 5$ products, whose intrinsic utilities $u_i(\theta_0)$ and $u_i(\theta_1)$ are randomly generated uniformly in $[0, 1]$ for all $i \in [n]$. For each random instance we run five different scenarios in which $\Lambda = k$ and $\text{Var}[\varepsilon] = k \pi^2 / (6 \mu^2)$ with $k = 10^\kappa$ for $\kappa = 0, 1, 2, 3, 4$. The rest of the parameters are kept fixed with $\mu = 1$, $r = 0.05$, and the terminal payoff function $G(\delta) = \max\{6 - 30\delta, 4 - 5\delta, 3\delta, -20 + 25\delta\}$. This is the same terminal payoff function that we used in the examples in Figures 2 and 4. Finally, in these and the rest of our numerical computations, we evaluate the value function of a given policy using Gauss–Seidel value iteration (see section 6.3 in Puterman 2005) with an error tolerance of 10^{-3} over a mesh of size 10^{-3} for the $[0, 1]$ interval that defines the domain of δ .

Table 3 presents the mean optimality gap—as well as the maximum value and standard deviation—

Table 3. Optimality Gap of the Asymptotic Policy

	$k = 1$	$k = 10$	$k = 100$	$k = 1,000$	$k = 10,000$
Mean, %	2.39	1.77	0.87	0.27	0.13
Maximum, %	26.19	22.95	13.67	1.99	0.87
Standard deviation, %	4.88	4.18	1.55	0.39	0.13

Note. Optimality Gap: $\Delta\Pi^A$.

computed over a run of 500 randomly generated instances. As we can see from the table, the two policies perform very well on average although the MV policy is substantially better than the A policy, especially for small value of k . As k grow large, both policies approach the optimal policy, which is consistent with our asymptotic analysis in Section 4. By comparing the “Maximum” rows that report the maximum optimality gap, we can also see that the MV policy is significantly more robust than the A policy for small values of k . The results in Tables 3 and 4 lead us to conclude that the MV heuristic dominates the A heuristic as it has consistently better optimality gap and comparable running times. For this reason, in the rest of our numerical experiments we focus exclusively on further exploring the performance the maximum volatility policy.

7.2. Benchmark Analysis

We next conduct a benchmark analysis in which we compare the performance of the MV policy against the following three alternative policies:

- Full Display (F): This policy always displays the entire set of prototypes, that is,

$$\text{Full Display Policy: } \mathcal{E}^F(\delta) := \{0, 1, 2, \dots, n\}. \quad (31)$$

Table 4. Optimality Gap of the Maximum Volatility Policy

	$k = 1$	$k = 10$	$k = 100$	$k = 1,000$	$k = 10,000$
Mean, %	0.56	0.12	0.26	0.15	0.09
Maximum, %	4.09	1.47	3.56	1.87	0.43
Standard deviation, %	0.86	0.22	0.41	0.22	0.06

Notes. Optimality Gap: $\Delta\Pi^{MV}$. Data: $\mu = 1, r = 0.05, G(\delta) = \max\{6 - 30\delta, 4 - 5\delta, 3\delta, -20 + 25\delta\}$, and $\Lambda = 2k, \text{Var}[\varepsilon] = k\pi^2/(6\mu^2)$.

This is a simple and popular benchmark that does not require any type of optimization.

- One-Step-Look-Ahead Policy Approximation (LA): This is a commonly used value function approximation, which, in our setting, corresponds to selecting an optimal experiment to display under the assumption that a final decision must be made after the outcome of this experiment is revealed. That is,

One-Step-Look-Ahead Policy :

$$\mathcal{E}^{LA}(\delta) \in \arg \max_{\mathcal{E} \in \mathcal{E}} \{\mathbb{E}_\delta[G(\delta + \eta(\delta, x, \mathcal{E}))]\}. \quad (32)$$

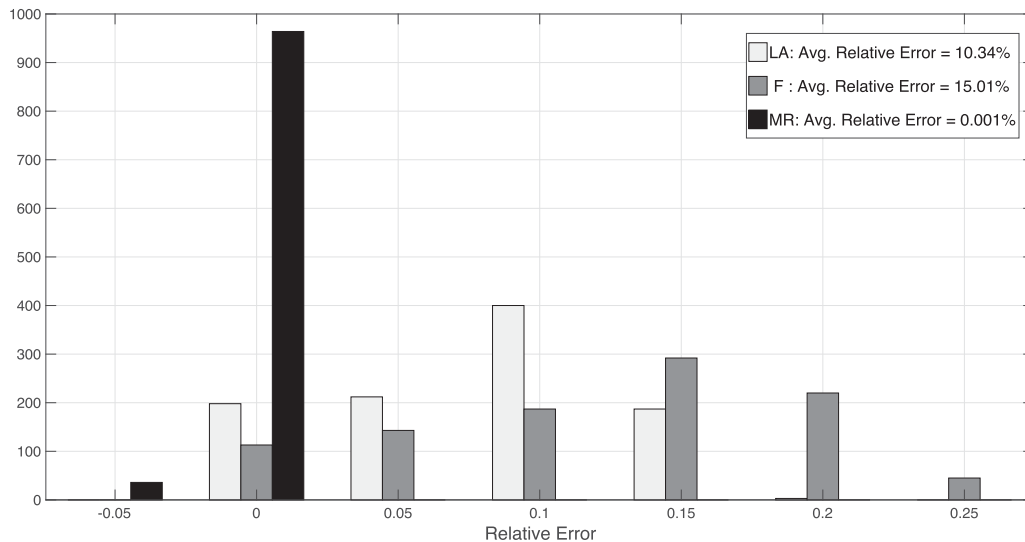
- Maximum Range Policy (MR): Motivated by the result in Proposition 10, we consider the policy

Maximum Range Policy:

$$\mathcal{E}^{MR}(\delta) := \arg \max_{0 \leq i \leq j \leq n} \mathbb{E}_0 \left[\frac{(1 - \mathcal{L}(\mathcal{E}[i, j]))^2}{\delta + (1 - \delta)\mathcal{L}(\mathcal{E}[i, j])} \right]. \quad (33)$$

A key advantage of the MR policy over the MV policy is that MR maximizes the instantaneous volatility

Figure 5. Distribution of Relative Error of the Full (F), One-Step-Look-Ahead (LA), and Maximum Range (MR) Policies Relative to the Maximum Volatility Policy over 1,000 Randomly Generated Instances



Note. Data: $\mu = 1, r = 0.05, G(\delta) = \max\{6 - 30\delta, 4 - 5\delta, 3\delta, -20 + 25\delta\}$, and $\Lambda = 2, \text{Var}[\varepsilon] = \pi^2/(6\mu^2)$.

over the smaller set of experiments $\mathcal{E}[i, j]$ defined in (28), which simplifies its computation.

We assess the performance of these three policies relative to the maximum volatility policy using the following relative error measure:

$$\text{Relative Error : } \bar{\Delta}\Pi^j = \int_0^1 \frac{\Pi^{\text{MV}}(\delta) - \Pi^j(\delta)}{\Pi^{\text{MV}}(\delta)} d\delta,$$

$j = \text{F, LA, MR.}$

Figure 5 shows the distribution of this relative error measure for 1,000 randomly generated instances of the problem with $n = 10$ products each.

As we can see from the figure, the MV policy substantially outperforms the LA and F policies, which have average relative errors of 10.34% and 15.01%, respectively. On the other hand, the MR policy is essentially equivalent to the MV policy with an average relative error of 0.001%. A similar conclusion holds when we compare the average running times of these policies. Indeed, the average running times per instance are equal to 0.292, 0.224, 0.520, and 16.184 for the MV, MR, F, and LA policies, respectively (all times in seconds).

In Online Appendix C, we confirm the superiority of our suggested policy by extending in various ways our numerical discussion. We extensively analyze and compare running times of different policies. We also look at the value of optimal stopping relative to the case in which the duration of the voting phase is predetermined. Finally, given the similarity with bandit problems, we compare our suggested MV policy to a couple of generic policies for MAB problems.

8. Conclusion

In this paper we study the problem faced by a DM who must select an action to maximize a reward function that depends on an unknown parameter that the DM can learn through experimentation. The DM has to decide dynamically which experiment to conduct and when to stop the experimentation. We adopt a novel diffusion-asymptotic analysis technique that relies on simultaneously increasing the frequency at which experiments are conducted and decreasing the degree of informativeness of each experiment. By doing so, the limiting experimentation process remains comparable to the “unscaled” process in terms of the informativeness per unit of time. This scaling fits many practical situations, specifically, online experimentation with which the velocity of data are always contrasted with their veracity. The diffusion model we obtain provides a number of important insights with respect to the nature of the problem and its solution. Interestingly, we also suggest a universal approach to unscale the diffusion approximation and derive a heuristic for the original

problem, that is, in the nonasymptotic regime, which is extremely relevant from an implementation point of view. Although the model that we study in this paper is very general, we have tested our solution approach in the context of an assortment-selection problem in which the experimentation is driven by consumers’ choices over menus of products. As a by-product of this analysis, we derive a diffusion limit for a belief process that is governed by an MNL model. Given the popularity of the MNL model to represent consumer preferences, we believe that our diffusion analysis and approximation have applications beyond the one discussed in this paper.

Our work opens also some interesting and natural research avenues. The diffusion approximation obtained by counterbalancing large sample sizes with little informativeness is revealed to be extremely effective and suggests that such approach should be considered in other related settings in which it might offer new approximations that complement those obtained by scaling only one parameter (e.g., the sample size). The recent arXiv manuscripts by Wager and Xu (2021) on MAB and Zenios and Wang (2021) in the context of clinical trials represent another confirmation of this claim. One ingredient of our model that made some of our analyses more tractable is the discrete set of experiments available to the DM from the start. In Section 5.1, we briefly discuss the design of the experimentation set in a way that leverages our model setup and analysis. We believe this is an interesting avenue to explore further by adopting probably a continuous and infinite set of possible experiments from which the DM selects dynamically the ones that are more effective for learning. Finally, our assumption that the unknown parameter takes only two values is restrictive, yet this assumption is an important first step in unravelling the multiple layers that the problem and the approach followed have to offer. The multihypothesis case requires a much more complex analysis and is left for a future work. We believe, for instance, that the principle of maximum volatility is preserved yet would require an adaptation of the definition of volatility to accommodate the multidimensional processes involved. As a result, the optimal experimentation becomes state dependent, and the diffusion limit of the belief processes requires even more advanced machinery than the one used in the two-hypothesis case. We also conjecture that the optimal stopping problem would again decouple from the dynamic experimentation. However, finding the stopping regions in this multidimensional setting will not be easy to characterize (see Dayanik et al. 2008) in the case of a compound Poisson process.

Acknowledgments

The authors are very grateful to the department editor, Omar Besbes, for encouraging them to further expand the scope of the proposed asymptotic regime in which the

information content of experiments is very low. The authors are also very grateful to the associate editor and three referees for their careful reading of the paper and for the many helpful and constructive comments.

Endnotes

¹ Making the wrong selection has even driven many major brands to discontinue some of their products shortly after introduction (see “Sell Big or Die Fast,” *New York Times*, J. Wortham and V.G. Kopytoff, August 23, 2011).

² We refer the reader to the *spot light* articles of the March–April 2020 issue of the *Harvard Business Review*.

³ This is a site on which anyone can design a T-shirt and submit it to a weekly contest. Viewers vote for their favorite T-shirts, the winning designs are selected for production, and their designers get rewarded.

⁴ To be precise, the probabilistic framework that we consider is defined by a probability space $(\Omega, \mathcal{F}; \mathbb{P}_0, \mathbb{P}_1)$ equipped with two probability measures \mathbb{P}_0 and \mathbb{P}_1 . For each $\delta \in [0, 1]$, we associate a probability measure $\mathbb{P}_\delta = \delta \mathbb{P}_0 + (1 - \delta) \mathbb{P}_1$ and let $\mathbb{E}_\delta[\cdot]$ denote its expectation operator. Finally, Θ is a Bernoulli random variable that satisfies $\mathbb{P}_\delta(\Theta = 0) = \delta$.

⁵ This description assumes that the outcomes of the experiments are instantly observed. Alternatively, we can think that each experiment takes an exponential random time (with rate Λ) to generate an outcome and only at this point in time can the next experiment be set. As a result, the outcomes of the experiments follow again a Poisson process with rate Λ . For instance, in the crowd-voting example mentioned in the introduction, experimentation occurs when a customer arrives to the online platform and votes, which we model as a Poisson process. See Section 6 for more details.

⁶ The notion that an experiment *dominates* another one is similar to the notion that an experiment is *more informative* than another one as discussed in Blackwell (1951) (see also Lindley 1956 and Le Cam 1996).

⁷ In Sections 5 and 6.2, we show how to interpret and operationalize our asymptotic scaling in practical settings.

⁸ During the voting phase, we assume that the arrival rate of voters Λ^k is $O(k)$, but during the selling phase, the arrival rate Λ_s^k does not need to be of the same order and could drop to $O(\sqrt{k})$. In this case, the different payoffs between the two hypotheses would still be significant.

References

- Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: A dynamic learning approach to assortment selection. *Oper. Res.* 67(5):1453–1485.
- Alaei S, Malekian A, Mostagir M (2016) A dynamic model of crowd-funding. Working paper, Ross School of Business, University of Michigan, Ann Arbor.
- Araman V, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5):1169–1188.
- Araman V, Caldentey R (2011) Revenue management with incomplete demand information. Cochran JJ, Cox LA, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley and Sons, Hoboken, NJ), 1–17.
- Armitage P, Berry G, Matthews JNS (2002) *Statistical Methods in Medical Research*, 4th ed. (Blackwell Science, MA).
- Bartroff J, Finkelman M, Lai TL (2008) Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika* 73(3):473–486.
- Bastani H, Bayati M, Khashayar K (2021) Mostly exploration-free algorithms for contextual bandits. *Management Sci.* 67(3):1329–1349.
- Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.
- Blackwell D (1951) Comparison of experiments. *Proc. Second Berkeley Sympos. Math. Statist. Probab.* (University of California Press, Berkeley, CA), 93–102.
- Blackwell D (1965) Discounted dynamic programming. *Ann. Math. Statist.* 36(1):226–235.
- Bolton P, Harris C (1999) Strategic experimentation. *Econometrica* 67(2):349–374.
- Breakwell J, Chernoff H (1964) Sequential tests for the mean of a normal distribution II (large t). *Ann. Math. Statist.* 35(1):162–173.
- Brezzi M, Lai TL (2002) Optimal learning and experimentation in bandit problems. *J. Econom. Dynam. Control* 27(1):87–108.
- Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.
- Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.
- Chang F, Lai TL (1987) Optimal stopping and dynamic allocation. *Adv. Appl. Probab.* 19(4):829–853.
- Chernoff H (1959) Sequential design of experiments. *Ann. Math. Statist.* 30(3):755–770.
- Chernoff H (1961) Sequential tests for the mean of a normal distribution. *Proc. Fourth Berkeley Sympos. Math. Statist. Probab.*, vol. 1 (University of California Press, Berkeley, CA), 612–624.
- Chernoff H (1972) *Sequential Analysis and Optimal Design* (SIAM, Philadelphia).
- Chick S, Frazier P (2012) Sequential sampling with economics of selection procedures. *Management Sci.* 58(3):550–569.
- Chick S, Gans N (2009) Economic analysis of simulation selection problems. *Management Sci.* 55(3):421–437.
- Dayanik S, Poor HV, Sezer SO (2008) Sequential multi-hypothesis testing for compound poisson processes. *Stochastics* 80(1):19–50.
- den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys Oper. Res. Management Sci.* 20(1):1–18.
- den Boer AV, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* 60(3):770–783.
- Fan L, Glynn PW (2021) Diffusion approximations for Thompson sampling. Preprint, submitted May 19, <https://arxiv.org/abs/2105.09232>.
- Feng Y, Caldentey R, Ryan CT (2021) Robust learning of consumer preferences. *Oper. Res.*, ePub ahead of print December 8, <https://doi.org/10.1287/opre.2021.2157>.
- Finkelman M (2008) On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *J. Ed. Behav. Statist.* 33(4):442–463.
- Gallego G, Talebian M (2012) Demand learning and dynamic pricing for multi-versions products. *J. Revenue Pricing Management* 11(3):303–318.
- Garivier A, Kaufmann E (2016) Optimal best arm identification with fixed confidence. *Proc. 29th Annual Conf. Learn. Theory*, vol. 49 (Columbia University, New York), 998–1027.
- Harrison JM, Sunar N (2015) Investment timing with incomplete information and multiple means of learning. *Oper. Res.* 63(2):442–457.
- Harrison JM, Keskin NB, Zeevi A (2012) Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Sci.* 58(3):570–586.
- Karatzas I, Shreve SE (1991) *Brownian Motion and Stochastic Calculus* (Springer-Verlag, New York).
- Kaufmann E, Cappé O, Garivier A (2016) On the complexity of best arm identification in multi-armed bandit models. *J. Machine Learn. Res.* 17(1):1–42.
- Keener R (1984) Second order efficiency in the sequential design of experiments. *Ann. Statist.* 12(2):510–532.
- Keskin G, Birge JR (2019) Dynamic selling mechanisms for product differentiation and learning. *Oper. Res.* 67(4):1069–1089.
- Keskin G, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* 6(5):1142–1167.

- Keskin G, Zeevi A (2018) On incomplete learning and certain-equivalence control. *Oper. Res.* 66(4):1136–1167.
- Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard Bus. Rev.*
- Kök AG, Fisher ML, Vaidyanatha R (2009) Assortment planning: Review of literature and industry practice. Agrawal N, Smith SA, eds. *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, International Series in Operations Research and Management Science (Springer, New York), 175–236.
- Lai TL (2001) Sequential analysis: Some classical problems and new challenges. *Statist. Sinica* 11(2):303–351.
- Le Cam L (1996) Comparison of experiments: A short review. Lecture Notes-Monograph Series, vol. 30 (Institute of Mathematical Statistics), 127–138.
- Lewis RA, Rao JR (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4):1941–1973.
- Lindley DV (1956) On a measure of the information provided by an experiment. *Ann. Math. Statist.* 27(4):986–1005.
- Marinesi S, Girotra K (2013) Information acquisition through customer voting systems. INSEAD Working Paper No. 2013/99/TOM, INSEAD, Fontainebleau, France.
- Naghshvar M, Javidi T (2013) Active sequential hypothesis testing. *Ann. Statist.* 41(6):2703–2738.
- Nahm M (2012) Data quality in clinical research. Richesson RL, Andrews JE, eds. *Clinical Research Informatics* (Springer, New York), 175–201.
- Oh M, Iyengar G (2019) Thompson sampling for multinomial logit contextual bandits. H. Wallach and H. Larochelle and A. Beygelzimer and F. d'Alché-Buc and E. Fox and R. Garnett, eds. *Adv. Neural Inform. Processing Systems*, vol. 32 (Curran Associates, Inc., Vancouver), 3145–3155.
- Papanastasiou Y, Bimpikis K, Savva N (2018) Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.
- Peskir G, Shiryaev AN (2006) *Optimal Stopping and Free-Boundary Problems* (Birkhäuser Verlag, Basel, Switzerland).
- Powell WB (2016) Perspectives of approximate dynamic programming. *Ann. Oper. Res.* 241:319–356.
- Puterman ML (2005) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).
- Qiu P (2014) *Introduction to Statistical Process Control* (Chapman & Hall/CRC, Boca Raton, FL).
- Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. (N.S.)* 58(5):527–535.
- Russo D (2020) Simple Bayesian algorithms for best arm identification. *Oper. Res.* 68(6):1625–1647.
- Sauré D, Zeevi A (2013) Optimal dynamic assortment planning with demand learning. *Manufacturing Service Oper. Management* 15(3):387–404.
- Siegmund D (1985) *Sequential Analysis: Tests and Confidence Intervals* (Springer-Verlag, New York).
- Soare M, Lazaric A, Munos R (2014) Best-Arm Identification in Linear Bandits. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems*, vol. 27 (Curran Associates, Inc., Montreal), 1–9.
- Sunar N, Birge JR, Vitavasiri S (2019) Optimal dynamic product development and launch for a network of customers. *Oper. Res.* 67(3):770–790.
- Ulu C, Honhon D, Alptekinoglu A (2012) Learning consumer tastes through dynamic assortments. *Oper. Res.* 60(4):833–849.
- Wager S, Xu K (2021) Diffusion asymptotics for sequential experiments. Preprint, submitted January 25, <https://arxiv.org/abs/2101.09855>.
- Wald A (1945) Sequential tests of statistical hypotheses. *Ann. Math. Statist.* 16(2):117–186.
- Wald A (1947) *Sequential Analysis* (John Wiley and Sons, New York).
- Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* 19(3):326–339.
- Zenios S, Wang Z (2021) Adaptive design of clinical trials: A sequential learning approach. Preprint, submitted January 27, <http://dx.doi.org/10.2139/ssrn.3713924>.