informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 42, No. 2, March–April 2023, pp. 271–292
ISSN 0732-2399 (print), ISSN 1526-548X (online)

A Multiarmed Bandit Approach for House Ads Recommendations

Nicolás Aramayo,[a] Mario Schiappacasse,[a] Marcel Goic[b,*]

[a] I2B Technologies, Santiago 7500010, Chile; [b] Department of Industrial Engineering, University of Chile, Santiago 8370456, Chile
* Corresponding author

Contact: nicoandres.aramayo@gmail.com (NA); mariochapav@gmail.com (MS); mgoic@dii.uchile.cl, https://orcid.org/0000-0002-9782-7447 (MG)

Received: July 17, 2020
Revised: November 29, 2021; April 21, 2022
Accepted: May 9, 2022
Published Online in Articles in Advance:
July 14, 2022

https://doi.org/10.1287/mksc.2022.1378

Abstract. Nowadays, websites use a variety of recommendation systems to decide the content to display to their visitors. In this work, we use a multiarmed bandit approach to dynamically select the combination of house ads to exhibit to a heterogeneous set of customers visiting the website of a large retailer. House ads correspond to promotional information displayed on the website to highlight some specific products and are an important marketing tool for online retailers. As the number of clicks they receive not only depends on their own attractiveness but also on how attractive are other products displayed around them, we decide about complete collections of ads that capture those interactions. Moreover, as ads can wear out, in our recommendations we allow for nonstationary rewards. Furthermore, considering the sparsity of customer-level information, we embed a deep neural network to provide personalized recommendations within a bandit scheme. We tested our methods in controlled experiments where we compared them against decisions made by an experienced team of managers and the recommendations of a variety of other bandit policies. Our results show a more active exploration of the decision space and a significant increment in click-through and add-to-cart rates.

History: Olivier Toubia served as the senior editor and Hema Yoganarasimhan served as associate editor for this article.

Funding: This work was supported by the Institute for Research in Market Imperfections and Public Policy [Grant ICM IS130002]. M. Schiappacasse gratefully acknowledges partial funding by the Institute for Research in Market Imperfections and Public Policy (IS130002 ANID).

Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mksc.2022.1378.

Keywords: multiarmed bandits • house ads • personalization • deep learning

1. Introduction

One of the most important challenges of the digital revolution is how to create and manage relevant content. Information sites such as digital newspapers and social network platforms are constantly applying a variety of methods to craft and select suitable content for their audiences. In retailing, this challenge translates into deciding an adequate marketing mix for each customer. For example, retailers must choose an attractive assortment, adequate promotions, and the right communication channel. Among all decisions that firms must make in their daily operations, we focus on the dynamic selection of house ads. House ads or internal links correspond to promotional information displayed on the retailer's website to highlight some specific products or category of products (Goic et al. 2018). House ads are an important component of the web design for most online retailers not only because they can have a direct impact in short-term sales but also because they can provide more consistent product offering that positively impacts customer satisfaction.

In recent years, many websites have started to use different recommendation systems to decide the content to display to their visitors, including association rules (Carmona et al. 2012), matrix factorization (Koren et al. 2009), and supervised learning (Agrawal et al. 2013). In this work, we address the problem of selecting a collection of house ads to exhibit in the homepage of a large retailer's transactional website to move customers forward in their purchase funnels. The problem is dynamic in nature for several reasons. For example, the set of available ads and the context in which those ads are displayed change over time. Furthermore, the effectiveness of a given ad can change over time because its attractiveness wears out (Braun and Moe 2013). A common practice in industry to deal with this dynamic problem is to decompose it in two stages: in an initial stage the decision maker evaluates the potential effectiveness of each ad by using randomized experimentation and then, in a second stage, marketers choose to display a combination of the best ads. However, it has been shown that this practice might be suboptimal because it does not explicitly consider the opportunity costs of the learning phase (Feit and Berman 2019).

A more efficient approach to deal with this problem is treating it as a multiarmed bandit (MAB) problem to define a strategy that minimizes the total expected regret (Vermorel and Mohri 2005, Villar et al. 2015). In a MAB framework, one must select a sequence of actions to maximize a cumulative reward with an imperfect knowledge of the performance of each action. By adopting this approach, we can explicitly address the well-known exploration-exploitation tradeoff, where we are simultaneously interested in learning what type of ads perform better and maximize the performance of the website by showing the best ads. Starting from the pioneering work from Robbins (1952), a variety of MAB policies have been proposed to deal with different problem settings (Kuleshov and Precup 2014). However, there are important features of the house ads problem that require us to deviate from the standard MAB policies. First, the identification of the most relevant content to show in the home page is customer-specific and therefore the set of house ads to display should depend on customer preferences. To accommodate this personalization, we make the rewards associated to each recommendation to be dependent on a vector of observable characteristics, which in the literature of MAB is commonly known as *contextual* bandits (Li et al. 2010). Our implementation of the context is based on a flexible mapping to connect customer characteristics to the rewards associated to each ad. This mapping not only allows us to provide personalized recommendations, but it also helps to learn about customer preferences with sparse transactional histories. In terms of the methodology, we connect customer level characteristics with personalized ads using *neural networks*. Despite losing some interpretability of intermediate results, we are mostly interested in providing a useful operational solution and therefore we believe this approach is instrumental for those purposes. The use of this technique requires us to depart from the standard supervised learning paradigm usually used to calibrate the underlying network and leads us to design a reinforcement learning version for the calibration. Reinforcement learning has been used in other domains (Mnih et al. 2013), but we are not aware of their use for training neural networks in the context of online advertising.

A second distinctive feature of our problem is its *combinatorial* nature. Unlike the traditional advertising display problem where a single slot is available to exhibit an ad (Schwartz et al. 2017), in our setting we need to decide an assortment of ads to be exhibited. By considering the whole list, we can incorporate business rules to define the combinations of ads that are feasible to be displayed simultaneously. More importantly, we can learn about what combinations are more effective. Although the existence of context effects are abundant in marketing literature (Simonson and Tversky 1992, Tversky and Simonson 1993, Chernev and Hamilton 2009), there is little work on practical methods to learn about those contextual effects with real size assortments. The approach we use, where we decide about the complete set to be displayed, captures direct interaction effects between ads. However, we do not have complete control over other elements on the website. Consider the case of a company running a national level campaign that is featured in the homepage. This campaign might have an impact in the performance of the house ads. For instance, if the retailer is running a campaign with a deep discount on mobile devices, a house ad promoting mobile accessories might have larger conversion rates. Other elements that can have an effect on the performance of a given ad include competition and the marketing mix in other channels (Verhoef et al. 2015). To account for these factors, we use *nonstationary* policies allowing for temporal variations in the rewards associated to each ad.

In summary, we developed a *contextual* MAB model with *flexible* learning to dynamically determine the list of house ads to display in the homepage of an electronic retailer to maximize the accumulated click-through rate of the ads. Although part of previous literature has focused on providing theoretical guarantees for simpler bandit policies, we use state-of-the-art tools to provide recommendations in the context of the House Ads problem. We perform a comprehensive simulation study to show that all the key components of our methodological proposal are useful to learn about the effectiveness of house ads and then we tested our methods in two controlled experiments where we compared them against decisions made by an experienced team of managers. Our results show that personalizing is associated with strong dominance in click-through and add-to-cart rates. Despite of focusing on the House Ads problem, several components of our model can be also applied to guide decision making in other settings beyond advertising. This is for example the case of pricing and assortment decisions that share several the key components of our methodological proposal, such as dynamic learning, personalization and cross-product effects.

The rest of the article is organized as follows. In Section 2, we review the relevant literature on online advertising and multiarmed bandit methods. In Section 3, we outline the main components of the decision problem and provide the primitives of our modeling approach. Next, we discuss how our model performs, first using simulated data (Section 3.3) and then in a field experiment with real display decisions (Section 4). In Section 5, we conduct a series of postexperimental evaluations to further illustrate the main drivers of the effectiveness of our bandit approach. We close with

Section 6 with final conclusions and directions for future research.

2. Literature Review

In this research, we use multiarmed bandits to dynamically decide a sequence of house ads to display in the homepage of an online retailer. We first review the methodological aspects of multiarmed bandit and then we discuss the literature on the substantive application.

From a methodological point of view, MABs are a widely used approach to solve the tension between the cost of acquisition of new information and the short-term benefits of using existing information. Although the foundations of MABs had been established a long time ago (Thompson 1933, Robbins 1952), in recent years, MABs have gain a renovated attention as a mechanism to provide feasible solutions to large stochastic dynamic problems. The algorithmic properties of MAB have been widely studied in the fields of statistics, computer science and operations research providing detailed characterizations of the optimality of different bandit policies for a variety of problem configurations including the case of combinatorial decisions (Chen et al. 2013, Ontanón 2017) and nonstationary rewards (Koulouriotis and Xanthopoulos 2008, Besbes et al. 2014) that we include in our model. Although previous research have considered these components in isolation, we are not aware of previous work that combines them all in a single MAB algorithm.

Regarding temporal evolution of rewards, although most of the MAB literature focuses on the stationary case, recent efforts have provided feasible solutions for the case of changing environments. For instance, Besbes et al. (2014) analyze the complexity of general class of problems with bounded nonstationary rewards. In our application, following the literature on the wear-out of online advertising (Chae et al. 2019), we expect the effectiveness of a given banner might decrease with more expositions, which is commonly known as *rotting bandits* (Seznec et al. 2019). In our model, we consider some of the ideas suggested in this literature such as using sliding windows (Levine et al. 2017) and give larger weights to more recent cases (Russac et al. 2019) but we adapt them to our policy that embeds a deep neural net.

Regarding the use of individual-level context data to provide personalized recommendations, with the exception a recent paper by Han et al. (2021), most of the literature on combinatorial bandits consider additive models. In these models, every time that an arm π is played, the algorithm receives information to learn about every combination that includes π, leading to important gains in the learning rates. However, a key feature of the House Ad problem is that the profit function is not additive. As is well known in the marketing literature, there are important interactions between ads that makes the click-through of a given banner dependent on the other ads shown. Although recent advances in combinatorial bandits using additive rewards could provide significant efficiency gains, they are not directly applicable to our case.

Considering the importance of banner interaction, our research is more closely related to the literature on dynamic assortment planning. This stream of research explicitly recognizes that the impact of adding an item to the recommendation critically depends on the other items in the set. For instance, Sauré and Zeevi (2013), study a stylized assortment problem and show that a simple *explore-first-and-exploit-later* policy can lead to a asymptotic regret bound of $O(K \log T)$. Closer to our work, Agrawal et al. (2019) propose an upper confidence bound type of algorithm that suggests assortments with better expected rewards but penalizes alternatives that have been more actively explored. For this policy, the authors show a worst-case regret of $O(\sqrt{KT \log KT})$. With respect to this literature, our work offers two important differences. First, unlike our work this literature relies on restrictive multinomial logit substitution patterns. Second, although this stream focuses on deriving theoretical properties and only provide numerical results with simulated data, our research focuses on describing the performance in a real setting with actual recommendations.

Recent literature on MAB has been active in studying practical applications in relevant business settings. For example, Bergemann and Välimäki (2002) use a bandit approach to decide optimal prices under uncertainty where players take into consideration the costs and benefits of learning. Bergemann and Hege (2005) developed a similar model to describe risky investments in innovations where investors and innovators cannot commit to future actions. Kleinberg and Leighton (2003) use bandits to model online auctions and discuss the impact of demand for information in the outcomes of the auctions. Other applications include dynamic assortment decision in the fast fashion industry (Caro and Gallien 2007) and network routing with uncertain delays (Awerbuch and Kleinberg 2004).

The application of MAB to support marketing decisions is scarcer. In an early work, Bertsimas and Mersereau (2007) developed a general framework to conduct adaptive experimentation in interactive marketing contexts, where decision makers can dynamically decide what is the most adequate message to deliver to their customers. Although they propose heuristics that can be implemented in real-sized problems, they only provide simulated results. The most common application of bandits in marketing contexts is the generation of dynamic recommendations for online display. For example, Li et al. (2010) use this framework to decide which article should be highlighted in a digital newspaper, and Tang et al. (2015) illustrate the application of MAB to personalize sponsored search advertising based on the keywords used. Except for a few exceptions, these applications are model free and use observed performance to learn from the reward structure of all available arms. Among those exceptions, we have Pandey et al. (2007) who aim to match online ads to websites. To reduce the complexity of the problem and to help gain more domain-specific insights, they use predefined taxonomies for ads and the web pages where ads could be shown. A more recent exception is the work by Schwartz et al. (2017) who propose a hierarchical model to capture how ad attributes can have an heterogeneous impact depending on the website characteristics. This study is similar to ours in that both test the performance of the proposed solution using controlled experiments. However, they decide which ad should be displayed in different websites, whereas we decide the collection of items to be displayed to each individual visitor in a single location.

Broadly speaking, the decision problem we are addressing in this research is related to the literature on display advertising which has been extensively studied in the past few years. Starting from Manchanda et al. (2006), who were one of the first in empirically demonstrating that exposure to online advertising leads to larger sales, several other investigations have discussed the effectiveness of different aspects of online advertising. For example, Bleier and Eisenbeiss (2015) analyze the interaction between retailers' trust and personalization depth in online advertising, and Kireyev et al. (2016) investigate the dynamic interaction between paid search and display ads. Regarding the conditions that favor more effective displays, Braun and Moe (2013) study how the different creatives of a single marketing campaign can have different impact on customer responses, whereas Breuer and Brettel (2012) analyze long-term impacts of banner advertising and other digital marketing tools.

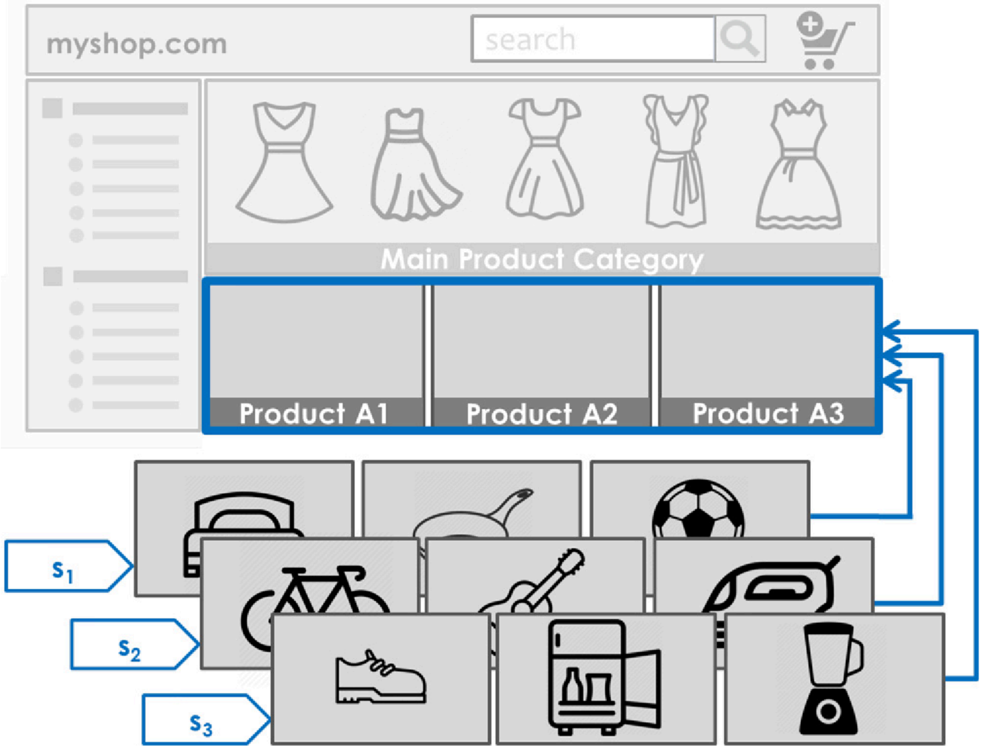
Our investigation focuses on house ads, which are a very specific type of online display. House ads are also called self-promotion ads or internal links and correspond to any promotional information presented on the retailer's website that is devoted to signal some specific products or categories of products. Unlike traditional online displays that are mostly devoted to increase traffic to the website, house ads are displayed to communicate specific elements of the value proposition to move customer forward in the purchase funnel. In terms of their execution, internal links also have important differences with traditional display advertising. For example, firms can decide the whole set of information surrounding internal ads. This gives more control to the retailer, but at the same time increases the complexity of the display decisions. Additionally, as the ads are displayed internally, there is no direct cost of advertising. Despite being considerably less studied than traditional displays, there are some investigations looking at their effects on consumer behavior. For example, using a Bayesian mixture approach, Rutz and Bucklin (2012) show that internal banners influence subsequent choices of page views during the current browsing session. More recently, using multivariate time series analysis, Goic et al. (2018) show that house ads have a direct impact in online sales, but a limited cross-channel effects. All these studies characterize the effect of internal ads on sales and are descriptive in nature. In our investigation we developed and tested the effectiveness of a prescriptive method to decide how these internal ads should be jointly displayed.

3. Problem Setup and Modeling Approach

Different online retailers have different structures to organize product information in their homepages. Common features include menus to explore product categories and search bars to look for specific information. The vast majority of homepages of online retailers include what we have named as house ads. This is a list of predefined slots where the firm can highlight some specific product information. Certainly, the content of the whole website can be decided using the methodology we propose here. However, most retailers only devote a certain area of the homepage to dynamically display house ads. The rest of the site is typically used for navigation tools or corporate-level information such as store openings, the introduction of new brands, or multichannel promotions. For the purpose of this research, we consider that the retailer has a fixed number of slots to allocate personalized information about specific products, and the content of rest of the website is decided elsewhere. An illustration of this structure is shown in Figure 1. In this figure, we show a schematic representation of the homepage where, in addition of search and navigation tools, there is a well-defined area to exhibit house ads. To do so, the firm can decide among a potentially large number of combination of ads. In this illustration we show three alternative sets of three ads (s_1, s_2, and s_3) that the retailer can decide to exhibit in the predefined area.

The selection of the combination of house ads to display for a visitor is frequently addressed by manual configuration of displays and constant monitoring of ads' performance over time by marketing teams and content managers. However, this problem can be formulated as a dynamic optimization problem. Formally speaking, consider a sequence of customers indexed by i ($i \in \{1, \ldots, I\}$) and a set of available ads

Figure 1. (Color online) Homepage with a Predefined Area for the Display of House Ads



indexed by a ($a \in \{1, \ldots, A\}$). The decision problem corresponds to the selection of a sequence of subsets to be shown to each visitor to maximize the expected reward associated to that sequence. However, not all combination of ads are admissible. For example, the subset must have a given length to accommodate the available space and it cannot have repeated items. Moreover, the set of feasible displays (or arms) can be further restricted to incorporate business rules designed to preserve long-term goals. For instance, the retailer might want to exhibit ads from different categories to provide more variety or can forbid the exhibition of two images next to each other because their aesthetic characteristics make them incompatible. In our model, we consider that any set of feasible ads (s_k) belongs to a set S that discards any undesirable combination of items. Although our prescriptive method can be extended for temporal variations in S associated to the addition of new ads or the deletion of others that have worn off, in our empirical evaluations we only consider scenarios where the set of feasible combinations is constant in the evaluation horizon. Thus, if w_{ik} is a decision variable taking the value of one if subset s_k is displayed to visitor i, then the underlying decision problem associated to the display of house ads can be expressed using the optimization problem displayed in Equation (1). For a similar formulation in a MAB context, see Schwartz et al. (2017, p. 504).

$$\max_{w} \pi = \mathbf{E}_y\left[\sum_{i=1}^{I}\sum_{k=1}^{K} w_{ik} \cdot y_{ikt}\right] \qquad (1)$$
$$\text{subject to} \sum_{k} w_{ik} = 1 \quad \forall i \in \{1, \ldots .I\}$$

In this formulation, y_{ikt} is the reward associated with displaying subset k to visitor i. As we will explain later, we have explicitly included a time index t to denote that rewards are nonstationary and that we learn from their values in batches ($t \in \{1, \ldots, T\}$). As the rewards are unknown, we take the expectation over their values. It is worth noting that w_{ik} must be decided sequentially and that the uncertainty on y_{ikt} depends on previous decisions. If a set k is more intensively shown, the uncertainty for that set would reduce. Thus, each period is different from the previous ones because there are different levels of knowledge about the rewards associated to each arm. This dependency introduces the key dynamic tradeoff of the problem where we aim to balance exploration (getting to know the performance of each super arm) and exploitation (displaying the best super arms). This problem could be solved using a dynamic programming approach (Bertsimas and Mersereau 2007), but it is very computationally demanding, especially in our case where the combinatorial nature of the ads to be displayed implies a potentially large number of alternatives (Powell 2007). Instead, in this research, we framed the problem as a multiarmed bandit that has been shown to provide workable solutions with good optimality conditions (Agrawal and Goyal 2012, Russo and Van Roy 2014).

In our multiarmed bandit formulation, each house ad corresponds to an arm. As we have argued, house ads should not be decided in isolation but as a combination of several ads. The combinatorial nature of the problem is referred in the literature as a combinatorial multiarmed bandit, where each feasible combination of arms of size N is defined as a *super arm*, $s_k \in S$ (Chen et al. 2013). This is a key feature of house ads decision, because the design of the whole website is under the control of the retailers, and therefore, they can decide combinations of ads that provide more variety (Kahn and Wansink 2004) or a better context to decide (Simonson and Tversky 1992). In our model, we allow that each banner combination have its own reward, which has the advantage of imposing no restriction on the substitution patters. Although learning about complete sets of banners works well in our application, it might not scale well for other settings with larger decision spaces.

In general, y_{ikt} can represent any perceived reward that the firm might be interested in lifting up; for example, the monetary value of the associated purchases, the length of the visitation session, or the probability the visitors return to the website. In our empirical applications, we will use click-through rates (CTRs), and later on, we also study the impact on add-to-cart (ATC) rates[1]; therefore, we define $y_{ikt} \in \{0,1\}\ \forall (i,k,t)$.

In addition to the combinatorial display of ads, the business context imposes a number of considerations that must be taken into account when selecting an adequate algorithm to solve the MAB problem. First, unlike most of the literature in MABs, we assume the rewards associated to the display of house ads are nonstationary. Our relaxation of this assumption is motivated because the context in which house ads are displayed changes over time. In fact, most online retailers include in their website seasonal information about brands or corporate promotions that can affect the relative attractiveness of any given ad. For example, if the retail chain is having special offer in the carpet category, any ad on that product category might be less attractive. Second, our method is designed to provide real-time recommendations to every customer visiting the website. However, in the practical implementation, the training of the model should be performed in batches, and the rewards are updated on a daily basis. This is why we include a time index t to denote that parameters can be updated periodically (Besbes et al. 2014, Schwartz et al. 2017). Third, for a large fraction of visitors, we have individual-level information that allow us for personalized recommendations. For example, we can observe the device they are using to visit (Goldstein and Hajaj 2022), the electronic channel they used to arrive to the website (Goić et al. 2022), or a history of previous website visitation (Park and Park 2016). Considering the sparsity of this individual-level information, we adopt a flexible approach using neural networks. The use of neural networks in the context of MABs is a distinctive feature of our proposed solution. Although in our empirical applications we use three layers in the network, the methodology is flexible enough to accommodate more layers to take further advantage of the recent developments in deep learning (Schmidhuber 2015).

Despite the significant progress in the identification of Internet visitors, for most electronic retailers, this is still a first-order concern, and they can only determine the identity of a fraction of them at the time of their arrival (Goic et al. 2021). To deal with this challenge, we implemented separated bandit policies depending on whether the customer can be identified at the time of arrival. The details of the algorithm to display personalized house ads for customers with individual-level information is discussed in Section 3.1. Considering the use of observed heterogeneity, we have labeled this algorithm as a contextual bandit. The algorithm to display ads for visitors with no identifiable link to purchase history is discussed in Section 3.2 that we have labeled as a noncontextual bandit.

3.1. Contextual Bandit

For cases in which we have information of the visitor at the time of the recommendation, we can approach the MAB problem as a contextual bandit (Li et al. 2010, Agrawal and Goyal 2013). In this context, we aim to design a policy that provides to each individual state (e.g., past purchase histories) a distribution of playing each super arm. The outcome of this policy is a personalized recommendation depending on the individual level data. Formally speaking, we look for a function that takes a state vector $\boldsymbol{x}$ and returns a recommendation of the set of ads to be displayed for that value of $\boldsymbol{x}$. In the state vector $\boldsymbol{x}$, we can include any information that might help determine which super arm is more likely to produce a positive outcome, such as demographics, past purchases, time, and device at the moment of navigation. A popular algorithm to solve this dynamic optimization problem under uncertainty is Thompson sampling (Thompson 1933). This principle states that it is optimal to select actions according to their probability of maximizing the expected reward at each decision stage. To use Thompson sampling, we will sample from the posterior distribution of each super arm and select the one with the highest probability of being the best action.

We use this principle and adapt it to the problem of house ad displays with individual level context.

To construct a flexible and accurate estimation of the probability that each combination of house ads provides the highest reward, we use a regularized feedforward *neural network* to sample from the posterior distribution of each super arm. Our selection of this method is justified not only because it provides a great deal of flexibility (Chen et al. 2005), but also because it has shown to perform well with an sparse set of individual level covariates $\boldsymbol{x}$ (Shepperd and Cartwright 2001). This is precisely the case we observe in purchase histories for website visitors. In fact, a typical customer only present a few purchases in a small number of product categories and no sales in many others. We consider a network with L layers, and we denote by $\boldsymbol{\theta}^l$ the vector of weights in the layer $l \in \{1, \ldots, L\}$. Thus, the output $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$ of the network depends on both the set of individual-level information $\boldsymbol{x}$ and the matrix of weights $\boldsymbol{\theta}$. The use of neural networks in the context of MAB is relatively scarce in the literature, and therefore we consider necesary to provide further details about a number of elements associated to the structure of the net and the learning mechanism we used to calibrate the underlying weights. In particular, we consider important to discuss the following three components:

1. **From supervised to reinforcement learning:** Neural networks are most commonly used in supervised learning, where a list of positive and negative labels are used to calibrate the internal weights of the network. In our problem, for each case we observe a positive or negative label only for the displayed ad, but there is no signal about the other ads. If we directly use a supervised learning paradigm to calibrate the net, the algorithm will myopically recommend only those ads that have relatively good performance, but it will underexplore those ads with no or little information.

Formally speaking, in supervised learning given a training example $\boldsymbol{x}$ and a target vector $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$, when training by backpropagation we match the estimation provided by each neuron in the output layer ($\hat{y_k}$) to $\boldsymbol{y}$ by using a multiclass cross-entropy cost function (Burges et al. 2007). In this paradigm, the entire output layer is trained using each positive or negative label. In contrast, in reinforcement learning, even if the reward is positive this does not imply that the other actions were incorrect. As we only observe errors with respect to the displayed action k, in the training procedure, we use a sigmoid cost function as follows:

$$J(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}) = -\sum_i \sum_k w_{ik}\big[y_k \cdot \log(\hat{y_k}(\theta_k)) + (1 - y_k) \cdot \log(1 - \hat{y_k}(\theta_k))\big]. \tag{2}$$

This function only affects the neurons associated to the displayed ads. The intuition behind this idea is that we cannot penalize for combinations that were not shown in a particular display. In supervised learning, we know whether the answer was correct or not, and therefore we use that information to train the entire output layer. In reinforcement learning, we do not know if showing another combination different than the one that was displayed could have provided a reward or not, so we only train one output neuron at a time and then back-propagate that error.

2. **Dropout to represent model uncertainty:** Neural networks in their most standard version produce deterministic outputs. Nevertheless, a key component of bandit policies is to randomize the optimal decision to balance the exploitation of the best solutions with the exploration of solutions with higher uncertainty. To implement this randomization, we use a Thompson sampling approach where we sample from the posterior distribution indicating the probability that each super arm is the the best action.

To do this, we turn to a regularization technique for neural networks named *dropout*. Dropout works by giving a zero weight to components of the input and hidden layers, with probability p, a hyper-parameter to be calibrated. We use dropout in training, which is a common practice in neural network model to prevent overfitting (Srivastava et al. 2014), but we also use it in the forecasting instance by randomly shutting off neurons in the network allowing us to represent model uncertainty and generating a distribution over the outputs. Then, we use Thompson sampling by selecting the highest estimated super arm according to the posterior distribution induced by dropout.

Certainly, there are other sampling schemes that can be implemented based on the output of the neural network. For instance, we could use an $\epsilon - greedy$ approach where we display the best superarm most of the time, but we randomly play suboptimal alternatives for an ϵ fraction of the cases. Our choice of using dropout to implement a Thompson sampling scheme is justified because previous literature suggest that, in the context of neural networks, dropouts can be used to characterize the posterior distribution of the forecast. This notion was first introduced by Gal and Ghahramani (2016), and it is further discussed in Online Appendix 1. In Online Appendix 2, we provide a simulation study to show how our proposal compares to alternative sampling schemes.

3. **Stochastic gradient descent for sequential learning:** We have argued that in our business context where we decide about house ads, it is important to explicitly consider the nonstationary nature of the outcomes. To accommodate this, we introduce an adaptation of the stochastic gradient descent (SGD) algorithm. In its most common form, SGD randomly select examples from the training set to approximate the updated direction of the network weights (Bottou 2010). In our

implementation, instead of randomly selecting examples from the training set, we consider the time where they occurred and therefore more recent case have larger incidence in calibrating the neural net. In this regard, we use stochastic gradient descent with momentum (Rumelhart et al. 1986), a variation of SGD that uses a moving average of the past m gradients where we give more importance to more recent gradients. Formally speaking, if v is the direction in which we update $\boldsymbol{\theta}$, then we update v according Equation (3):

$$v \leftarrow \alpha_1 v - \alpha_2 \frac{1}{m} \nabla_\theta \sum_{i=1}^{m} J(\boldsymbol{\theta} \mid \boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), \tag{3}$$

where $\alpha_1 \in [0,1)$ and $\alpha_2 > 0$ control for the exponential decay and learning rates respectively. This way, the latest changes in the environment can provide the direction in the learning process before we produce a recommendation. This is similar to the idea proposed in Sutton and Barto (2018) and Russac et al. (2019), who modeled the MAB problem with nonstationary reward distributions by giving more weight to recent observations.

Taking all these components into consideration, the resulting algorithm we used to make recommendations for visitors with individual level information is presented in Algorithm 1. Here we assume that the set of feasible superarms S is already defined and the algorithm is trained on a data set $\mathcal{D}$, where each context $\boldsymbol{x}^{(i)}$ is paired with the corresponding vector $\boldsymbol{y}^{(i)}$ that contains the reward of each of the K feasible combinations. As we combine Thompson sampling with deep neural networks, we also refer to this algorithm as *deep-Thompson*.

Algorithm 1 (Deep-Thompson Sampling)

```
Set $\alpha_1, \alpha_2, p \in [0,1]$, the number of hidden layers $L$ and neurons per layer $N^l$
Initialize weights $\boldsymbol{\theta}$ of the network $\boldsymbol{h}(\cdot)$
for each $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in \mathcal{D}$ do
    $h^{(0)} = \boldsymbol{x}$
    for $l$ in $1, \ldots, L$ do
        for $n$ in $1, \ldots, N^l$ do
            $\theta^{l,n} = \begin{cases} \theta^{l,n} & \text{w.p. } (1-p) \\ 0 & \text{w.p. } p \end{cases}$
        end
        $h^{(l)} = f(\boldsymbol{\theta}^l h^{l-1})$
    end for
    $\hat{\boldsymbol{y}} = \mathrm{MNL}(h^{(L)})$
    Compute the loss function $J(\boldsymbol{\theta})$
    Apply the back-propagation algorithm with gradient descent with momentum using $(\alpha_1, \alpha_2)$ and $J(\boldsymbol{\theta})$
end for
To predict the superarm to display given a state vector $\boldsymbol{x}^{(n)}$, pass it through the trained network $\boldsymbol{h}(\cdot)$ with dropout and select the superarm $k$ with the highest estimated probability of success, $\hat{y}_k$.
```

In the algorithm, $f(\cdot)$ denotes the activation function used to determine whether a given signal is tresspassed to the next layer. As is usual practice in neural networks, in our application we use rectified linear unit (ReLU) activation functions. To initialize the weights of the network we used He initialization (He et al. 2015). The idea behind this procedure is to make the variance of the output of a layer equal to the variance of its inputs to help convergence of models with multiple layers. To speed convergence up, we calibrated the mean values of the output layer to reproduce historical CTR at the aggregated level and then use ReLU for all other weights in the net.

The implemented network had three layers, and in the last layer, we include a multinomial logit model (MNL) to define the estimated reward $\hat{\boldsymbol{y}}$. Once the model is estimated, for each new case characterized by a vector $\boldsymbol{x}^{(n)}$, we apply random dropout and execute a single stochastic forward pass through the network, selecting the alternative with the highest probability of success.

In our empirical application, the dropout parameter was set at $p = 0.1$ so there is a 10% chance of any neuron to turn off in the input or hidden layers. Finally, the SGD with momentum parameters was set at $(\alpha_1, \alpha_2) = (0.1, 0.01)$. These are conservative choices frequently used in other applications. The network is trained by performing back-propagation one training example at a time from the set of all observed cases $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ in the historical data set $\mathcal{D}$. Last, considering the stochastic gradient of Equation (3), the data are fed in a time-ordered manner to take advantage of the properties of SGD with momentum.

3.2. Noncontextual Bandit

In our main model, we make personalized recommendations depending on customers characteristics. However, not all visitors can be identified at the time of their arrival, and the practical implementation requires to also provide recommendation for them. From a methodological perspective, we are mostly interested in analyzing the performance of the full model of Section 3.1. However, this noncontextual scenario allow us to understand how our bandit policies compares against simple recommendations made to nonidentified users that are likely to visit less frequently and being less familiarized with the home page of the retailer.

For simplicity, we produce noncontextual recommendations based on a batched ϵ-greedy approach that we adapted to accommodate nonstationary rewards. The ϵ-greedy algorithm selects the action with the highest estimate for the mean reward an ϵ proportion of time while selecting the rest of actions at random. Thus, this algorithm plays greedily an $(1-\epsilon)\%$ of the time and learns on the performance of other super arms in the remaining fraction. In this setting, it is desirable to start actively exploring the solution space and progressively reduce this exploration as we collect more information. To do so, we use an exponential decay for ϵ_t ($\epsilon_t = (|S| \cdot t \log(t))^{-\delta}$).

The decay parameter δ controls for the rate at which we diminish the exploration probability. A value of δ closer to one makes these probabilities approach zero at a faster rate. This ϵ_t scheme provides an upper bound on the expected regret (difference in expectation with respect to an oracle that always plays the best action) at each time step (Slivkins 2019). To accommodate nonstationary rewards, we estimate the expected reward of a super arm using an exponential recency-weighted average, as in Sutton and Barto (2018). Despite of its simplicity, ϵ-greedy algorithms are shown to perform well in practice (Schwartz et al. 2017, Riquelme et al. 2018). The resulting version of our nonstationary ϵ_t-greedy algorithm is formally described in Algorithm 2.

Algorithm 2 (Batched Nonstationary ϵ_t-Greedy)

```
Set $\alpha, \delta \in [0,1]$, $\epsilon_1 = 0$ and $\hat{y}_{k1} = 0 \ \forall k$
for $t \in 1, \ldots, T$ do
    $\epsilon_t = (|S| \cdot t \log(t))^{-\delta}$
    for $i \in 1, \ldots, |I_t|$ do
        Sample $c \sim U[0,1]$
        if $c > \epsilon_t$ then
            Choose a super arm $k$ randomly from $S$ and observe reward $y_{ikt}$
        else
            Choose $\text{argmax}_{k \in S} \hat{y}_{kt}$ and observe $y_{ikt}$
        end if
    end for
    $\hat{y}_{kt} = \frac{1}{|I_t|} \sum_i y_{ikt}$
    $\hat{y}_{k,t+1} = (1-\alpha) \cdot \hat{y}_{kt} + \alpha \cdot \hat{y}_{k,t+1}$
end for
```

In Algorithm 2, I_t is the set of visitors that arrived during day t. Also, $\hat{y}_{kt}$ is the success rate of super arm k up to day t. In the calibration of $\hat{y}_{kt}$, α controls for the weight we give to present information, and $(1-\alpha)$ is the corresponding weight for historic rewards. A value of α closer to one represents a strong belief that the reward distributions change rapidly on this environment.

3.3. Simulation Study

To evaluate the performance of our main bandit policy, we ran a simulation study where recommendations were made over synthetic data. The objective of this study is to demonstrate that our model has the potential of outperforming other policies and evaluate how different components of our proposal contributes to increase click-trough rates. This simulation follows the idea in van Emden and Kaptein (2018) for evaluating contextual multiarmed bandit algorithms. Here, we consider a binary reward $y_{ia(k)} \in \{0,1\}$ for displaying an ad a, within a combination of ads s_k to visitor i. We model the reward structure according to a standard multinomial logit model:

$$\Pr(y_{ia(k)} = 1 \,|\, \boldsymbol{x_i}, \alpha_k, \beta_k, \gamma_k, \lambda_{a\cdot}) = \frac{\exp(\alpha_k + \boldsymbol{x_i'}\beta_k + \gamma_k \cdot n_{kt} + \sum_{b \in s_k} \lambda_{ab})}{1 + \sum_{a' \in s_k} \exp(\alpha_k + x_i'\beta_k + \gamma_k \cdot n_{kt} + \sum_{b \in s_k} \lambda_{a'b})}. \tag{4}$$

The vector $\boldsymbol{x_i}$ correspond to individual level information defining the context, and β_k captures the relationship between that context with the reward structure. The parameter α_k is a fixed effect devoted to capture that some banners are consistently more effective than others. Next, n_{kt} is the number of times the arm k has been displayed and it varies dynamically according to the recommendations of each algorithm. Consequently, the parameter γ_k is meant to capture nonstationary rewards. Finally, the set of parameters λ_{ab} captures the interaction of simultaneously displaying ads a and b, such that if $\lambda_{ab} < 0$, banner a decreases the probability of being clicked when jointly displayed with banner b. To build the simulated scenario, we assumed that $x_i \sim N(0,1)$. The rest of the parameters are also generated from normal distributions with α_k chosen to have similar CTR to what we observe in a realistic setting. We tried other distributions, and the simulation results remain qualitatively unaltered.

These simulations can be used to compare our model against any arbitrary recommendation policy. However, for simplicity, we will focus on evaluating the most relevant features of our bandit algorithm. More specifically, we use the simulated data to produce recommendations with the following algorithms:

- **Deep-Thompson:** This is the full model we propose to use in the live experiment, and it considers all the key components we have described as critical for deciding displays of house ads. This includes nonstationary rewards and contextual recommendations using an embedded deep network.
- **Top-k:** This model is identical to the full model, but we estimate the reward of each banner separately with no interactions ($\lambda_{ab} = 0$). Here, the actual recommendation corresponds to the three banners with the highest reward. By ignoring the potential interactions between banners, this allow us to assess the impact of not considering the combinatorial nature of the decision.
- **Stationary Rewards:** This model is identical to the full model, but before feeding the model, we shuffle the order of the cases in the training batch, and instead of giving more importance to more recent cases, all considered cases weight the same. Thus, this model

allow us to assess the impact of ignoring that rewards might be nonstationary.

- **No Context:** This model is identical to the full model, but instead of calibrating a neural network that uses the context as input, we completely ignore the contexts and estimate homogeneous reward for each banner combination. In the absence of context, we update the reward of each alternative using a standard Beta-Bernoulli distribution (Chapelle and Li 2011).
- **Thompson MNL:** This model is identical to the full model, but instead of using a deep network to describe the relationship between context and recommendations, we link purchase histories using a simple multinomial logit. This formulation has been used in previous applications (Agrawal et al. 2020, Oh and Iyengar 2021). In particular, this is the same formulation proposed by Schwartz et al. (2017), but without the hierarchical structure as we only consider one website. In this synthetic evaluation, the data are generated using exactly the same functional form we assume in this policy, and therefore, it is expected that this algorithm can lead to unrealistically good performance that would not replicate in live settings. In the experiment with real data, this model will provide a benchmark to evaluate the gains of using a complex nonlinear transformations to connect context to recommendations.
- **Oracle:** This policy can anticipate the reward of each superarm with no uncertainty, and therefore, the oracle knows the probability of each action and thus can always choose the superarm with the highest probability. Although this policy is not implementable with actual decisions, it is useful to generate an upper bound of what can be achieved using a bandit policy.
- **Random:** In this model, we simple recommend a random combination of ads. This model is simply used to provide a lower bound to the click-through rates and enable us to have an assessment of in which degree the model can lift cumulative rewards.

By varying the matrix of contexts and the corresponding weights, we generated 100 scenarios that we used to evaluate the performance of all algorithms. A summary of these results is presented in Table 1, where we display mean and standard deviations of the cumulative rewards in these 100 scenarios.

These numbers provide preliminary evidence that the proposed deep-Thompson model performs well, and it leads to higher click-through than most of the benchmarks we analyzed. As we anticipated, the Thompson-MNL is the only (feasible) model that exhibits better perfomance. This is because the synthetic data are generated using exactly the same functional form of this model. In practice, however, this functional form is unknown, and we expect that allowing for more flexible relations should lead to better recommendations. Overall, these simulations demonstrate that the proposed algorithm has the potential of generating more profitable recommendations. However, we do not know in which extent banner interactions, stationary rewards, and the complexity of the relationship between context and recommendations materialize in an online setting. This is formally evaluated in Section 5.1, where we use the experimental data to conduct off-policy evaluations. In particular, we will show that with real decisions our model outperforms all other benchmarks, including the Thompson-MNL.

Table 1. Mean and Standard Deviation of Click-Through Rates of Different Bandit Policies on Simulated Data

Bandit policy	Mean	Standard deviation
Oracle	0.337	0.129
Deep Thompson	0.260	0.114
Thompson MNL	0.310	0.129
Stationary rewards	0.239	0.103
No context	0.216	0.096
Top-k	0.140	0.068
Random	0.078	0.035

4. Experimental Evaluation

In this project, we partnered with a regional retailer who competes in the department store market. The retail chain operates a few dozens brick and mortar stores in four countries and has a fully transactional website that accounts for around 10% of its total sales. To evaluate the impact using of multiarmed bandit algorithms to dynamically select the house ads that are displayed, we ran two experiments where our automatic recommendation engine was compared against an alternative benchmark. In the first experiment, a set of six banners was made available from which three must be displayed for every visitor. We compared the performance of our recommendation algorithms in contextual and noncontextual settings against a fixed display chosen by an experienced marketing team from the retailer. In the second experiment, the decision was made from a set of eight available banners, and we used a more challenging benchmark. In this case, the marketing team can adjust the display configuration at any time by looking at the intermediate results of the bandit algorithm.

To calibrate the model, the retailer gave us access to historical purchases in the online channel in the last two years. For confidentially reasons, we cannot disclose the exact number of customers purchasing in different categories, but in the data set, we observe more than half a million unique customers that made more than 2 million purchases from more than 200,000 different SKUs. Frequency of purchases are relatively small, and about 40% made more than two purchases in these two years (about 16% made more than five purchases). Thus, as is common in online retailing, the associated customer-product incidence matrix is extremely sparse with a density of 0.0016% nonzero cells.

To alleviate the sparsity of the data, the aforementioned list of products is classified in 26 large product categories. Figure 2 shows the purchased items distribution by these product categories[2] and illustrate that even after aggregating to a category level, there is a large variation between categories and several of them exhibit very small purchase incidences. This is another reason why we use a flexible mapping to connect individual level covariates and display decisions.

Next, we explain how we implemented the experiments and discuss some key results to show how our proposed methodology largely outperforms the current practices of the firm.

4.1. Experiment 1

In the first experiment, for each visitor, we must select three house ads to display from a set of six possible ads. As we decide about combinations of ads, our decision space consist of a total of 20 feasible superarms, from a total of 120 possible combinations. In this set, we discarded combinations with repeated ads, and we make no distinction between the order of the ads by excluding permutations of the same collections of ads. The ads consist of women's fashion, makeup, personal beauty accessories, home appliances, toys, and linens. It is worth noting that the ads with beauty accessories and the ones with home appliances both explicit advertised discounts. As we explicitly consider two cases (with and without context), this first experiment consisted of four experimental conditions as illustrated in Figure 3. In simple words, we start by classifying visitors depending on the availability of purchase histories. Customers with previous purchases are candidates to be recommended with our contextual bandit policy. Here, we randomly selected half of identifiable visitors to be treated with our recommendations, and the other half is assigned to the control group. Similarly, customers without purchase history are randomized to be recommended to our noncontextual policy or a control. In this design, we decide the bandit policy based on whether we can identify the user and not in the length of the purchase history. Thus, if the users are identified, they receive a recommendation from the contextual bandit regardless of how many times they have visited or purchased in the past.

Regardless of the use of context, in our experimental design, we use the *business as usual* practice as the experimental controls. This is letting a group of experienced marketing managers select the combination of ads they believe can have the best performance. Our experiment was carried out in a three-week period, from September 11 to October 3, 2018, and the combination of house ads recommended by the group of experts does not change over time (in the next experiment, we relax this restriction). In this period, more than 50,000 visitors participated in the analysis. In this experiment, 33.4% of customers were identified at the time of arrival having previous purchase histories, whereas the remaining 66.6% had no data associated with them. As a result, more than 9,000 visitors were assigned to one of the two conditions of the contextual experiment, and more than 19,000 were assigned to each condition in the noncontextual one.

Figure 2. Distribution of Purchased Items by Product Category

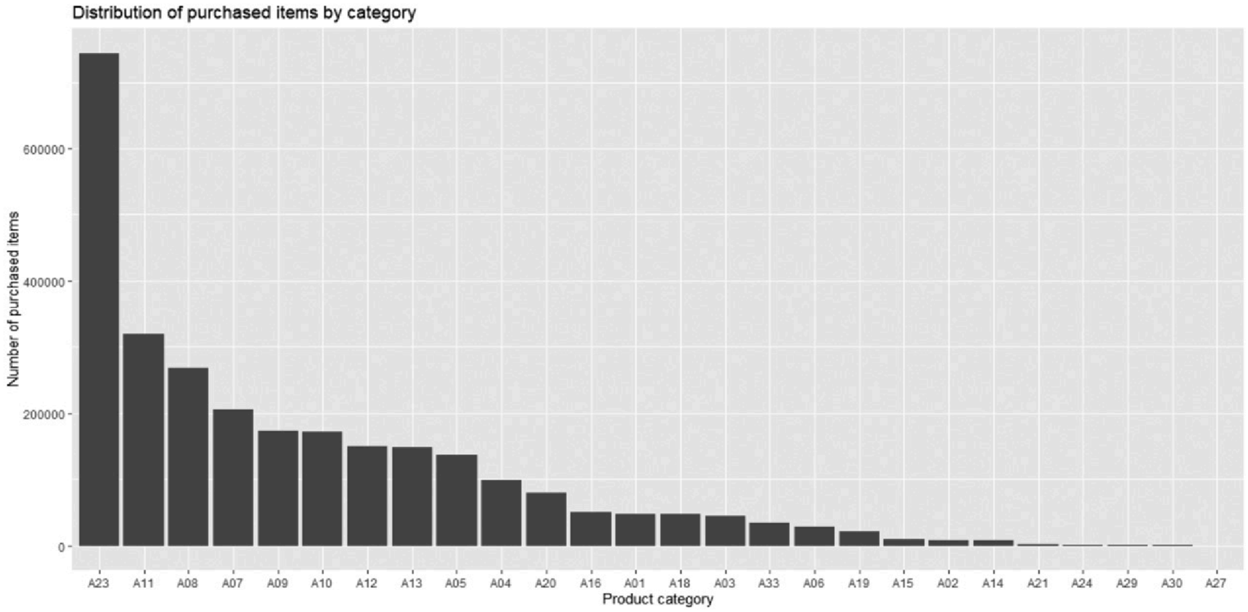
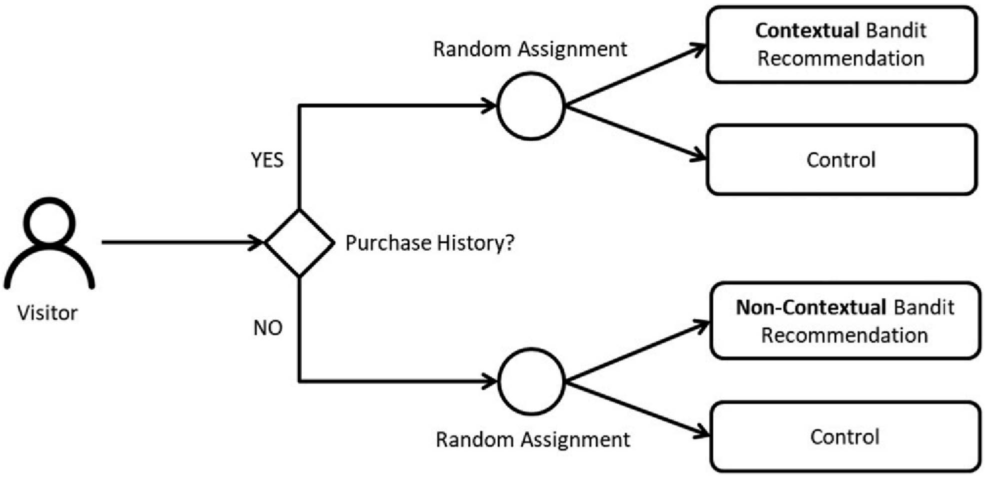


Figure 3. Experimental Assignment

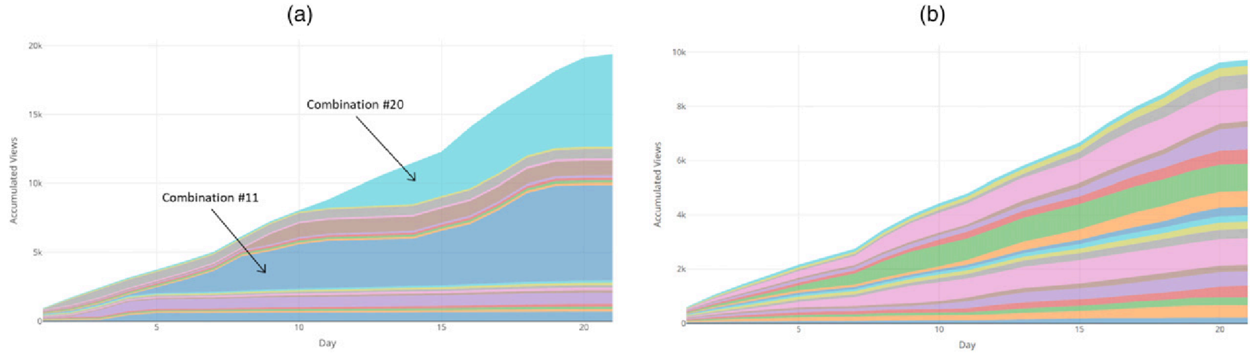


In Figure 4, (a) and (b), we show a graphic representation of the solution profiles of both bandit solutions. As expected, the noncontextual solution starts with a fairly homogeneous distribution of exhibitions among available combinations, but after a few days, the algorithm starts exploiting superarm 11. However, even after this initial exploitation, the algorithm leaves room for exploration. In fact, after a week, the algorithm also starts exploiting superarm 20. The solution profile of the contextual bandit is similar in that it does not exploit any solution in the first few days, but it departs from noncontextual bandits in several ways. To start with, even in early phases of exploitation, the algorithm does not concentrate their solutions in a single superarm, but instead heavily plays three of them. This is a direct consequence of personalization, and it implies the identification of three latent segments of customers. In later stages of the process, the algorithm has more information to connect customer profiles with a given combination of ads, and therefore, the resulting solution is much more diverse depending on customer characteristics. In fact, at the end of the evaluation period, there are a few superarms that concentrate most of the displays, but all arms were played with some probability. To see formal evaluations of Gini indices reporting the variability in each each bandit policy for the distribution of the exhibition of different ads, see Section 3 in the online appendix.

A critical question of the experimental evaluation is whether these recommendations do generate larger rewards. Recall that for this experiment, our goal is to maximize the accumulated click-though rates. Table 2 precisely reports aggregated lift for CTR for both policies. These results indicate that both bandit models achieved superior rewards, and the difference is highly statistically significant for the corresponding difference in the proportions test. Moreover, the improvement in both cases is larger than 15%, which is managerially relevant.

To better understand the dynamics behind the aggregated lift in CTR, Figure 5, (a) and (b), displays the cumulative CTR for both bandit policies against

Figure 4. (Color online) Accumulated Displays of Each Combination for Bandit Recommendations



Notes. (a) Solution for the noncontextual case. (b) Solution for the contextual bandit.

Table 2. Summary Results of Experiment 1 in Noncontextual and Contextual Settings Against a Fixed Human Decision

	Contextual		Noncontextual	
	Lift	*p* value	Lift	*p* value
CTR	15.9%	0.03	15.4%	< 0.01

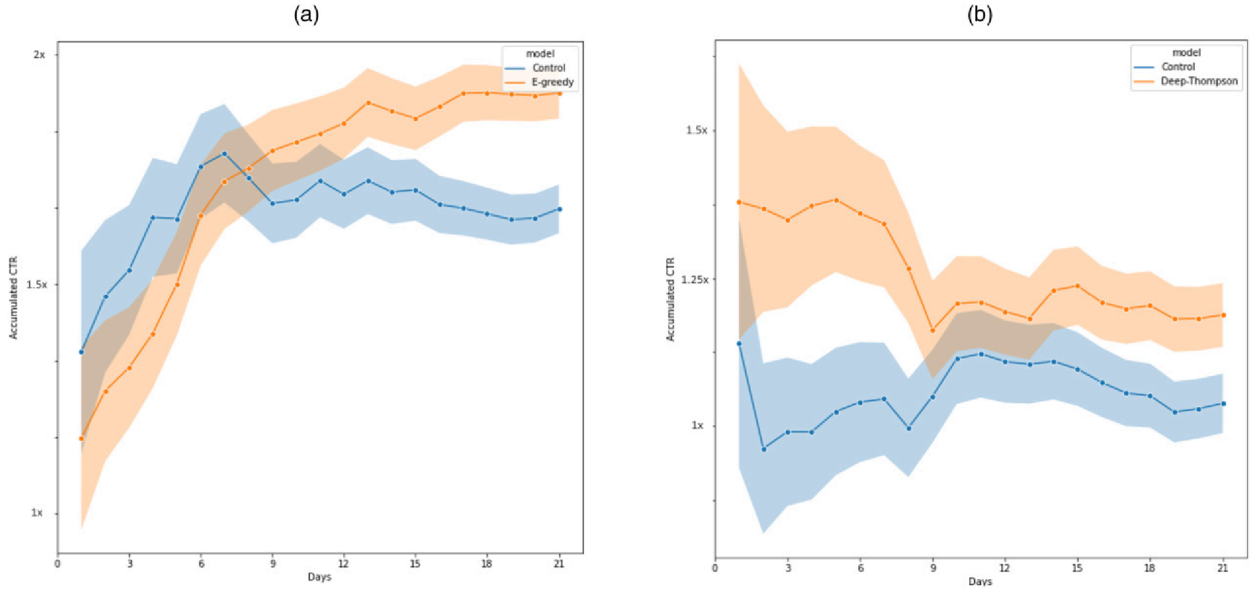
their corresponding controls. The actual CTRs have been masked from the plot for confidentiality, but they all range between 1% and 10%. Interestingly, the trajectory of both policies is quite different, which can be explained not only by the algorithm itself, but also because they provide recommendations for different customer groups. For example, the solution provided by experts outperformed the noncontextual bandit for the first couple of days. This indicates that the marketing team has enough business knowledge to identify a solution that leads to better than average click-through. On the other hand, the cumulative CTR of the contextual bandit is above the expert curve over the whole evaluation period. We believe this is a good signal that the customer data are informative to personalize the recommendations and that the several thousands of ads we display every day provides enough information for the MAB algorithms to learn fast.

When comparing the curves, we also observe that for the case of noncontextual bandit, the cumulative CTR appears to be increasing over time reflecting that the algorithm properly learns and exploits. However, this is not the case for the contextual bandit. We hypothesize that this is mostly caused by a wear-out effect, in which customers find the ads less attractive over time. The reason why wear-outs are mostly present in the second case is because this group includes more recurrent customers that are exposed to the ads multiple times. We further explore this issue in Section 5.2.

4.2. Experiment 2

The basic setting for this second experiment remains, but there are a number of differences that make the problem more challenging. First, we have more house ads available. In this second experiment, we had a set of 56 feasible superarms (resulting from eight individual house ads) that is almost three times larger than in the first experiment. Furthermore, the experiment only ran for 13 days, which is about half of the time we used previously. In consequence, we have less time to learn about the effectiveness of a larger set of ad combinations. Second, the marketing team was allowed to change their original combination at any time to simulate their usual form of operation. It is worth noting that the marketing team can observe not only the performance of their policy, but also the performance of our MAB strategy, and therefore they can exploit the exploration capabilities of our method. The decision of making the intermediate results of MAB available to the marketing teams is based on the configuration of the analytic platform used by our partner firm that simultaneously displays key performance indexes of all live conditions. Third, the location where house ads are displayed in the website

Figure 5. (Color online) Accumulated CTR of ϵ-Greedy Algorithm (Left) and Deep-Thompson Sampler (Right) with Respect to the Fixed Display During Experiment 1



Notes. (a) Noncontextual bandit. (b) Contextual bandit.

moved slightly upward, and therefore we had access to a larger number of views per day. Finally, and taking into consideration the results of the first experiment, we only used 20% of the visitors as controls, resulting in more than 35,000 visitors in the bandit treatment and more than 8,000 in the control. The ad categories in this second experiment were personal beauty, tableware, linens, and video games—all four house ads with explicit discounts of up to 60%, 60%, 70%, and 30%, respectively—electric scooters, electric tools, cameras, and watches, which offer no explicit discounts.

In terms of the methodology, we made two adjustments. First, in this second experiment, we focused exclusively in the main model. This is justified not only because it contains the key features we consider relevant for the recommendations of House Ads, but also because the retailer we partnered with was primarily interested in evaluating the impact of providing personalized recommendations. Second, in addition to CTRs, we complement our evaluation with a stronger conversion metric. For technical reasons, the visitor sessions could only be tracked up to the checkout page, and therefore, we do not observe if customers actually complete the purchase or not. However, we do observe if customers added a product to the shopping cart, and we use this behavior to build add-to-cart (ATC) rates that we also include in the evaluation. Formally speaking, we built this metric as the fraction of customers who ended up adding a product to the shopping cart from those available in the landing page associated to the corresponding house ad before his current navigation session expires. It is worth noting that, despite tracking ATC, the bandit algorithm is still trained using CTR as rewards.

Table 3 shows the key performance results of this second experiment. As in the first experiment, the MAB approach leads to much larger CTR achieving a 36.8% gain with respect to the team of experts. The comparison with respect to ATC is more challenging not only because the ATC rates are smaller, but also because we take no further interventions in the rest of the conversion funnel. Thus, any positive effect in ATC can be interpreted as a better identification of customer preferences. As is shown in the table, the resulting difference is outstanding with an increment of more than 99% in ATC rates. In Online Appendix 4, we provide additional analysis to support that ATC rates' improvement can be explained not only for a better selection of ads, but also for a more personalized recommendation.

Table 3. Summary Results of Experiment 2 Against Expert Decision Making

	Contextual	
	Lift	p value
CTR	65.12%	< 0.01
ATC	99.34%	< 0.01

To shed further light on the dynamics of these gains, we report in Figure 6(a) the accumulated CTR and in Figure 6(b) the ATC. As in the first experiment, the bandit solution generates better CTR from the very early stages. Interestingly, the same pattern spills over to ATC. It is interesting to see that the solution by the team of experts presented an important improvement in the last four days. This is because they decided to change their house ads configuration to only exhibit the superarm with the largest CTR. This indicates that even if the retailer does not want to fully automate, they can benefit from having an efficient algorithm to detect the most effective combination of house ads. The positive short-term impact of this myopic change does not necessarily imply an effective policy in the midterm. This is because it inhibits learning from new ads and because it does not adapt to wear out effects.

In the next section, we discuss how the observed performance compares with the recommendation of alternative policies, and we further explore the reason why the proposed policy leads to better recommendations.

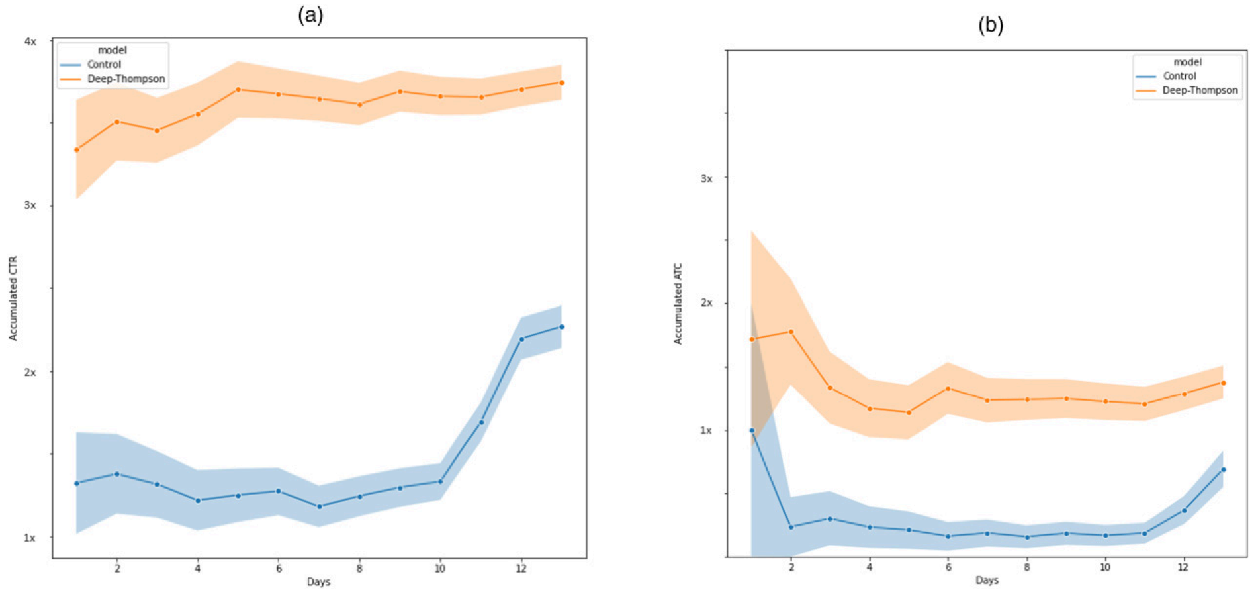
5. Postexperimental Evaluation

5.1. Off-Policy Evaluations

Our experimental results indicate that the recommendations derived from our *deep-Thompson* policy lead to large improvements in key business metrics with respect to current practices of the firm. In addition, In the simulation study of Section 3.3, we demonstrate that our proposed policy was capable to capture contextual data and provide superior recommendations with nonstationary rewards. However, thus far we have only compared our solution to more sophisticated benchmarks with synthetic data and not with actual customer responses.

A direct way for making these comparisons is by implementing the full list of alternative algorithms and comparing their online performance. The main limitation of using that approach in our setting is that we are working with the home page of a single retailer and therefore the evaluation of multiple algorithms will necessary imply splitting the traffic for different recommendation systems limiting our statistical power. Fortunately, recent work on bandit learning has shown that off-policy evaluation can provide inexpensive and fast comparison of different algorithms, because they can be applied with historical data that were collected from a different recommendation policy (Agarwal et al. 2017). The basic intuition of this offline evaluations is that we can identify proper events in the logged data that can be used to build *valid* counterfactual scenarios.

Figure 6. (Color online) Accumulated Performance of the Deep-Thompson Sampler Algorithm in Click-Through Rates (Left) and Add-to-Cart Rates (Right) with Respect to the Dynamic Display During Experiment 2



Notes. (a) Click-through rates (CTR). (b) Add to cart rates (ATC).

To evaluate the performance of a policy $\mathcal{P}$ using offline data, we use the method proposed by Li et al. (2011). Here, we observe the historical sequence of observed plays $(x_{it}, w_{ik}, y_{ikt})$ and, if it happens that the policy $\mathcal{P}$ makes the same decision w_{ik} suggested by the logging policy, the event is retained and used to compute the payoff. If the policy $\mathcal{P}$ recommends a different set of banners, the event is ignored and we proceed to the next event. The algorithm resembles a rejection sampler (Gilks and Wild 1992), and it leverages the large number of plays we observe to generate meaningful counterfactual scenarios. More recent developments have proposed other methods to conduct off-policy evaluations that might be more computationally efficient (Swaminathan and Joachims 2015, Thomas and Brunskill 2016, Agarwal et al. 2017), but this is not a critical concern in our case. We use off-policy to compare our bandit policy against the same series of benchmarks we used in the evaluation with synthetic data in Section 3.3, except for the *oracle* policy that cannot be computed without the explicit definition of the data generating process. Each of these benchmarks is devoted to evaluate the different features of our proposal. For instance, the *Thompson MNL* model is useful to evaluate if having a complex deep neural provides a benefit with respect to a simpler multinomial logit. Similarly, the model with *stationary rewards* allow us to assess if using stochastic gradient descent with momentum to compute the learning direction is indeed preferable to a model where we assume that CTRs are stable over time.

In the counterfactual evaluation of each of these benchmark policies, every period we sample 30,000 bandit events from the logged data and keep those for which the recommendation played in the live experiment coincides withe the policy recommendation. This procedure might lead to different sample sizes between different recommendation algorithms. Thus, to compare different policies on samples of the same size, we randomly select a subset of events of each policy. Moreover, to make sure the evaluation does not depend on the sample, we simulated each benchmark policy 300 times and report the mean reward and the corresponding standard deviations. Results of these off-policy evaluations are summarized in Table 4.

Results of these off-policy evaluations across all scenarios indicate that the *deep-Thompson* algorithm we propose is indeed the policy that leads to the highest mean reward, providing strong support to the idea that all key modelling components we include in the model are indeed useful to generate better recommendations. Although the performance of the *Thompson-*

Table 4. Mean and Standard Deviation of Off-Policy Simulation of Different Bandit Policies

Bandit policy	Mean	Standard deviation
Deep Thompson	0.2078	0.0257
Thompson MNL	0.1903	0.0809
Stationary rewards	0.1021	0.0283
No context	0.0960	0.0040
Top-k	0.0620	0.0069
Random	0.0454	0.0018

MNL closely follows the performance of the full model, our *deep-Thompson* approach largely outperform all other policies. For instance, our recommendations generate more than three times more clicks than the *Top-k* strategy that myopically chooses single banners ignoring that their performance can depend on the other ads displayed. Considering these interactions are important for the design of attractive displays; in the next section, we further explore the nature of these cross-banner interactions.

The algorithm also provides large improvements with respect to the model with *no context* and the one with *stationary rewards* nearly doubling their CTRs. The former result simply reinforces the importance of aligning the marketing mix with individual preferences (Montgomery and Smith 2009) and that previous purchases have large predictive power in explaining future behavior (Rossi et al. 1996). The latter comparison indicates that in this setting, the CTRs effectively vary over time and that our bandit policy calibrated using stochastic gradient descent with momentum is effective in capturing these variations. Again, considering the importance of advertising dynamics in the marketing literature (Chen et al. 2016, Chae et al. 2019), in the next section we provide further analysis to understand how the effectiveness of the ads vary over time. Beyond mean rewards, the relative ordering in performance is similar if we look at the frequency in which each algorithm generates the best mean reward across scenarios, where we find that our proposed model provides the largest cumulative reward in 57.6% of the cases, whereas the *Thompson-MNL* version leads to more clicks in 41.3% of the scenarios (the remaining 0.1% correspond to the model with *stationary rewards*).

The good performance of the *Thompson-MNL* algorithm invites further discussion of its results. One interpretation for this relatively high number of clicks generated by this model is that the underlying linear utility function is a good approximation of the actual relation between context and rewards. We pose that if the retailer only has limited data, a multinomial logit model could suffice to generate adequate recommendations. However, the availability of larger and more complicated historical data at the individual level should favor better perfomance of machine learning models. Our evaluation shows that, for this case, in terms of the magnitude, compared with the *Thompson-MNL*, the proposed model already generates an additional 1.75 percentage points, which represent a 9.19% improvement in CTRs. We expect this gain should be even larger if, in addition of purchase in different categories, we add frequency and recency of purchases in different categories, whether they purchase on promotion, and other transactional information commonly available for omnichannel retailers (Goic and Olivares 2019). The proposed algorithm not only exhibits better CTRs, but it also leads to more robust solutions. In fact, the *deep-Thompson* exhibits much more stable rewards between scenarios with a reduction of three times in the standard deviation of the rewards compared with its closest contestant. In summary, despite the good performance of the *Thompson-MNL*, our results show that allowing for a deep-neural mapping between context and rewards, we can attain meaningful improvements in performance.

Beyond the aggregated performance statistics, the evolution of the rewards provides additional insights about how each algorithm learns dynamically. The evolution of mean rewards for all recommendations policies are displayed in Figure 7. According to this figure, deep-Thompson, Thompson MNL, and the model with stationary rewards learn relatively quickly, exhibiting relatively high performance in the first five days. However, the latter policy start decreasing its effectiveness as it has no room to correct for temporal variations of banners' rewards over time.

It is also worth noting that the MNL models start learning faster than the more flexible model with a deep network. However, after a few days of training, the more flexible deep-Thompson keeps improving, whereas the rewards of the MNL-based models gets flatter. We hypothesize that if the model only has a short horizon for learning, a MNL model can provide an efficient mechanism to connect the context to the rewards. However, the disposition of longer training periods allows our proposed model to further learn bout more complex relations that are not allowed in the fixed functional form imposed by the multinomial logit.

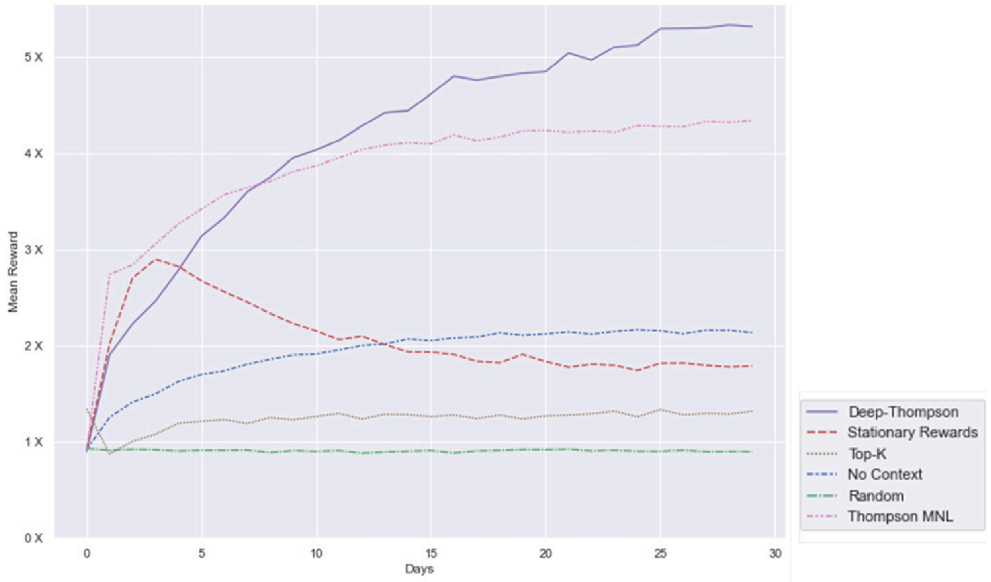
Regarding the policies that ignore the context (No-Context) and the existence of banner interactions (Top-k), in Table 4, we already documented they have sizable worst perfomance than the proposed model. From Figure 7, we can visualize that halfway to the evaluation horizon, both policies reached their best performance. This provides further indication that banner interactions and context play important roles in House-Ads recommendations, and therefore, managers aiming to implement automatic display decisions need to consider both components in their policies.

Taking all these results together, we conclude that the proposed model leads to the best performance among all policies we evaluated. Although the Thompson-MNL model could provide a good approximation especially in the case with limited data, in our setting, our proposal leads to higher and more stable click-through.

5.2. Drivers of MAB Effectiveness

In previous sections, we demonstrated that our recommendations provide significant gains with respect to expert recommendations and alternative recommendation policies. In this section, we analyze the nature of recommendations played by the bandit

Figure 7. (Color online) Evolution of Mean Reward for Different Bandit Policies



algorithms to better understand why they lead to better performance.

Two central premises that guide our modeling efforts are that the attractiveness of internal ads depends on what other ads were displayed around them and that the reward structure is nonstationary. Based on the CTRs, we can verify if these assumptions are indeed relevant in the context of house ads decisions. As our recommendations are based on a neural network, we have no explicit knowledge about why some banner combinations lead to better performance. After observing the actual recommendations of the algorithms, we can analyze those responses to understand why our MAB algorithm led to better recommendations. In this section, we explore the nature of banner interactions, we investigate the existence of advertising wear-out, and we explore the role of users' purchase histories in the effectiveness of our MAB policy.

Formally speaking, let u_{ijt} be a latent utility that visitor i experiences for clicking a banner j displayed on day t. To corroborate that rewards are nonstationary and the existence of banner interactions, we define u_{ijt} as follows:

$$u_{ijt} = \lambda_j^0 + \lambda_j^w n_{jt} + \sum_{h \neq j} \lambda_h^c \delta_{ihj} + \lambda^p TP_i + \varepsilon_{ijt}. \qquad (5)$$

In this specification, λ_j^0 are banner fixed effects and capture that some ads are more attractive than others. In the experiment, we display combinations of ads, and therefore, every ad is displayed accompanied by a variety of other ads. Consequently, there is no need to omit the fixed effect of any given ad. The variable n_{jt} is the number of times that the ad j has been displayed up to day t, and therefore λ_j^w capture how the attractiveness of each ad wears out. There might be other dynamic elements that can lead to nonstationary rewards, but in this analysis, we only consider previous exhibitions because we believe it drives most of the temporal variation. In the model, we also have binary indicators δ_{ihj} that take the value of one if banner j was jointly shown with banner h to visitor i. Thus, parameters λ_h^c capture any interaction that banner h can have when displayed with other banners. To complete the model, we have another binary variable TP_i that takes the value of one if the combination of ads shown to visitor i include two banners announcing price discounts. Considering that some of the available banners explicitly indicated deep price promotions (e.g., 40% off), we include this variable to see how promotions further influence customer attention. To estimate the model, we assume that ε_{ijt} are extreme value distributed, and we use the data collected in experiment 1 to calibrate a binary logit regression. In this model, the dependent variable corresponds to binary indicators to denote if a given visitor clicked in a given ad.

Results of this model are displayed in Table 5. These results are very consistent with our previous description of the problem of house ads. First, by looking at banner fixed effects, we find that some ads are more attractive than others, and therefore, it is worth exploring those differences, taking advantage of the efficiency of a bandit policy. Second, we also find that the attractiveness of a given ad varies over time. In fact, half of the ads exhibit significant wear-

Table 5. Logistic Regression Analysis for Wearout Effects and Promotions

Coefficient	Estimate	Significance
Fixed effect house ad 1	−4.49	***
Fixed effect house ad 2	−4.34	***
Fixed effect house ad 3	−2.99	***
Fixed effect house ad 4	−2.85	***
Fixed effect house ad 5	−3.91	***
Fixed effect house ad 6	−4.57	***
Wear out effect house ad 1	−1.4e-4	**
Wear out effect house ad 2	3.3e-7	
Wear out effect house ad 3	−6.3e-6	***
Wear out effect house ad 4	4.6e-6	
Wear out effect house ad 5	−4.0e-5	***
Wear out effect house ad 6	1.0e-5	***
Interaction with house ad 1	−0.66	***
Interaction with house ad 2	−0.51	***
Interaction with house ad 3	0.80	***
Interaction with house ad 4	0.92	***
Interaction with house ad 5	−0.12	
Interaction with house ad 6	−0.76	***
Two discounts	−1.62	***

***, **, and * Significance levels: 99.9%, 99%, and 95%, respectively.

out effects; that is, we found evidence that as these banners are shown to more visitors, the likelihood of being clicked decreases significantly. One of the ads exhibits a positive dynamic coefficient, implying that it got more attractive over time. This effect is probably explained by different factors not associated to wear-outs, but it also validates the importance of modeling a nonstationary reward.

Parameter estimates on interaction are also supportive of our modeling approach. In fact, we find that five of the six ads create consistent interactions. If a given banner is jointly displayed with 1, 2, or 6, it has a lower probability of being clicked, whereas having the company of banners 3 and 4 increases the likelihood of being clicked. The negative effect of displaying two house ads with explicit discounts also supports the existence of interactions. In simple terms, this negative coefficient indicates that a superarm that simultaneously displays more than one ad with a discount underperforms compared with other combinations that only include one banner displaying price cuts. Interestingly, the bandit policies did capture most of these regularities. For example, superarm 20, which was heavily displayed by our policies, consists of house ads 4, 5, and 6, which is precisely the combination of ads with the highest sum of fixed effects among combinations with at most one price discounts.[3]

Another important question is if our contextual bandit algorithm can effectively take advantage of customer-level information. In the simulation study, we demonstrate the the algorithm can recover individual level differences from an underlying linear utility model. However, this is not guaranteed in real problems where the relationship between purchase histories and browsing behavior can be more complicated. In the actual MAB policy, we use a neural net that can efficiently deal with large vector of attributes. Here, as we are interested in explainability, we summarize purchase history of visitor i in two simple metrics: Q_i is the total number of products the visitor has purchased from the retailer in the last two years, and D_i is the number of different categories that the customer has purchased in the same time frame. In this summary, Q_i measures the quantity, and D_i measures the diversity in the purchase history. Using these metrics, we rely again in a binary logistic regression model to describe the likelihood that a visitor i clicks in any of the recommended ads. The underlying utility function is given in Equation (6).

$$v_i = \mu_0 + (\mu^{mq} Q_i + \mu^{md} D_i) MAB_i \\ + (\mu^{cq} Q_i + \mu^{cd} D_i)(1 - MAB_i) + \xi_i \qquad (6)$$

In this model, MAB_i indicates whether visitor i was assigned to the treatment. Thus, parameters μ^{mq} and μ^{cq} capture how larger quantities in the purchase history correlate with larger CTRs in treatment and controls. Similarly, μ^{md} and μ^{cd} capture how more product diversity in purchase histories affect conversion in treatment and controls. Results of this regression are reported in Table 6.

These results indicate that diversity of previous purchases is a strong indicator that a customer will click in a house ad, but quantity is actually inversely correlated with click-through. To see if our personalized recommendation can take advantage of individual level data, we are interested in the difference of the coefficients between MAB policies and the control. Here we found that the MAB coefficients are always larger than those associated to the controls, but the difference is only significant for the metric of diversity ($p = 0.03$). This indicates that when a customer has purchased a wider variety of products, a contextual bandit algorithm provides better house ads recommendations than the control. Nonetheless, the total number of products does not produce better results by itself. This can be explained because the frequency of purchases that might be associated to other

Table 6. Logistic Regression Analysis for Personalization Effects

Coefficient	Estimate	Significance
Intercept	−3.140	***
Quantity × control	−0.008	*
Diversity × control	0.070	***
Quantity × MAB	−0.006	**
Diversity × MAB	0.103	***

***, **, and * Significance levels: 99.9%, 99%, and 95%, respectively.

demographic factors and because a visitor who only purchased in one or two categories is less likely to have any history in categories that are related to the products being promoted in the set of available house ads.

6. Conclusions

In this study, we are concerned with how to decide the display of house ads in the home page of an online retailer. As the rewards associated to the display of each ad are uncertain, the use of MAB policies appears as an efficient approach to balance the dynamic learning of the effectiveness of each item and the short-term exhibition of the most profitable ads. Moreover, as we know individual level information for a relevant fraction of the customers, the algorithm can be extended to provide personalized recommendations for each customer profile. The combination of bandit policies with flexible learning to provide personalized recommendations is novel in the marketing literature, and it requires the use of a number of complementary techniques to capture all relevant features of the house ad problem. To start with, the number of clicks received by an internal ad does not only depends on their own attractiveness but also on how attractive other products displayed around them are. Thus, to decide about a complete collection of ads that captures those interactions, we use a combinatorial type of MAB algorithm. In our implementation, instead of using the standard supervised learning paradigm to train the underlying neural network that maps individual-level data with actions, we use a reinforcement learning criteria that avoids getting stuck in local optima. Furthermore, we allow for nonstationary rewards by giving more importance to recent displays in the training of both bandit algorithms.

We tested our implementation of MAB to recommend house ads in two field experiments. This way, we compared the performance of our solution against the recommendations of an internal team of experts by randomly assigning visitors at their arrival to one of those conditions. Our results indicate that our methodology significantly outperformed the baseline and that the improvements are relevant from a business point of view. Moreover, in the second experiment, we also tracked ATC rates, and they were also significantly improved with our recommendations. This provides a stronger signal that our system not only provides more engagement for customers but that a better fit with customer preferences can be monetized for larger profits. To complement these aggregated results, we compared the profile of actions generated by our policy to the baseline. In addition to the larger ability to systematically explore the decision space, the bandit policies provides more diverse set of solutions. More importantly, it captures relevant cross-effects between ads, which provides a strong justification to decide about combinations of items instead of individual ads. In our analysis, we also included a comprehensive comparison of the performance of our proposed bandit policy, and we found that each of the key components of the model contributes to higher CTRs. Based on this result, we pose that managers interested in implementing a MAB approach to decide personalized House Ads recommendations should consider the interaction of banner combinations and nonstationary rewards. Furthermore, if extensive individual data are available, the use of deep learning to map context into recommendations could lead to an additional lift in performance.

We believe our proposed model is a step forward in the design of personalized experiences in online retailing. However, there are a number of promising avenues for future research. First, in our analysis, we decide the combination of ads to display while keeping all other elements of the homepage unaltered. Future research can relax these constraints to decide most of the content in an integrated model. In addition, the analysis of the structure of the website can shed a more comprehensive understanding of how different components of the website affects conversions. For example, following Tang et al. (2013), we could expand the use MAB to learn about the most effective layouts to display the personalized content. Second, in our empirical evaluation, we assume customer rewards are binary (e.g., click-through). In our case, this is justified because, in real time, we do not observe whether the customer actually purchased an item or not. If that information was available, we could use a continuous reward structure (Bertsimas and Mersereau 2007) to prioritize those ads, leading to more profitable product categories. Third, in terms of the characteristics of the ads, we only consider if they announce price discounts. However, previous research shows that other factors such as size and spatial location of ads can play an important role in advertising effectiveness (Marszałkowski and Drozdowski 2013, Goic et al. 2018). Although we have little variation in the data to inform about the impact of these factors, they could enrich the decision space in other bandit applications.

The algorithm we proposed is suitable to incorporate dynamic changes in the set of available combinations of ads (S) by solving the problem in batches. However, in the experimental studies, our methods were only tested with a constant set of ads. Given the nature of our expert recommendation baseline that is slow in learning about new context, we expect that dynamic variations in S can lead to an even larger difference with respect to the control. If the firm can dynamically decide the duration of each arm, the methodology could be further developed incorporating *mortal* bandits (Chakrabarti et al. 2008). Nowadays, firms use fixed calendars for the exhibition of each the ads. Using real-time performance to also decide the deletion of underperforming banners is a promising avenue for future research. In our analysis, we used a limited set of customer level information to produce personalized recommendations. The expansion of the set of covariates could lead to even better results and to a more personalized navigation experience. For example, the results of experiment 1 provide some evidence that the effectiveness of house ads can wear out, especially for returning customers. In the set of covariates we used, we did not consider if the customer had been exposed before, but this could be integrated with no significant change in the algorithm itself. By adding this information, the recommendations can be personalized not only between visitors, but also between navigation sessions. Furthermore, in our investigation we compare our policy against an extensive number of alternative bandit policies. However, this is an active area of research with numerous alternative algorithm that could provide further improvements to our recommendations. This is, for example, the case of regression oracles (Foster and Rakhlin 2020) and the translation from reinforcement learning to offline regression (Simchi-Levi and Xu 2022).

In terms of scalability, the proposed methodology should accommodate the addition of more features and more customers with no additional burden. However, because of its combinatorial nature, the use of larger sets of ads could be challenging. Although in our two empirical studies, we consider relatively small decision spaces of 20 and 56 superarms, in other applications, marketers might be interested in deciding about a larger number of alternatives. We numerically simulated how the computational time required to estimate the model and make recommendations grows for larger decision spaces and, even for a decision space of 1,000 superarms, those times only increase marginally. Although these are promising figures, further research is needed to corroborate them in a live setting. In this regard, the design of efficient algorithms to scale combinatorial bandits is still an area of active research (Wen et al. 2015, Wang et al. 2017).

To conclude, in this study we implemented two separated versions of bandit algorithms depending on our ability to identify visitors at their arrival. We believe that the contextual model can be extended to generate recommendations to any visitor. In fact, even for customers that enter anonymously or those who have no previous records with the company, the retailer observe a number of covariates that can be used to provide some level of personalization. For example, the device they use to visit the website (e.g., desktop computer or mobile device) and the channel they used to visit (e.g., direct load, organic search) or even the time at which they arrive. Thus, our contextual model can be used for all customers, but for some of them, the personalization is based on a full set of covariates that use historical purchases and for others we only used contextual information about how they arrived to the site.

Endnotes

[1] Actual purchases are observed but not in real time and therefore they cannot be used to train the model in practice. This is because purchases are processed by a different system that manages the payment validation and made logistic decisions.

[2] As the exact number of sales per categories cannot be disclosed, in this illustration we multiply the series by a random number in [0.8, 1.2].

[3] There are other combinations such as ads 3, 4, and 5 with a larger sum of fixed effects, but this combination includes two discounts and therefore, to derive its mean value we would need to discount the coefficient $\lambda^p = -1.62$.

References

Agarwal A, Basu S, Schnabel T, Joachims T (2017) Effective evaluation using logged bandit feedback from multiple loggers. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 687–696.

Agrawal P, Avadhanula V, Tulabandhula T (2020) A tractable online learning algorithm for the multinomial logit contextual bandit. Preprint, submitted March 7, https://arxiv.org/abs/2011.14033.

Agrawal R, Gupta A, Prabhu Y, Varma M (2013) Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. *Proc. 22nd Internat. Conf. World Wide Web*, 13–24.

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *Proc. Conf. on Learn. Theory*, 39–41.

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *Proc. Internat. Conf. on Machine Learn.*, 127–135.

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

Awerbuch B, Kleinberg RD (2004) Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. *Proc. 36th Annual ACM Sympos. on Theory of Comput.* (ACM, New York), 45–53.

Bergemann D, Hege U (2005) The financing of innovation: Learning and stopping. *RAND J. Econom.* 36(4):719–752.

Bergemann D, Välimäki J (2002) Information acquisition and efficient mechanism design. *Econometrica* 70(3):1007–1033.

Bertsimas D, Mersereau AJ (2007) A learning approach for interactive marketing to a customer segment. *Oper. Res.* 55(6):1120–1135.

Besbes O, Gur Y, Zeevi A (2014) Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 199–207.

Bleier A, Eisenbeiss M (2015) The importance of trust for personalized online advertising. *J. Retailing* 91(3):390–409.

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. *Proc. COMPSTAT* (Springer, Berlin), 177–186.

Braun M, Moe WW (2013) Online display advertising: Modeling the effects of multiple creatives and individual impression histories. *Marketing Sci.* 32(5):753–767.

Breuer R, Brettel M (2012) Short-and long-term effects of online advertising: Differences between new and existing customers. *J. Interactive Marketing* 26(3):155–166.

Burges CJ, Ragno R, Le QV (2007) Learning to rank with nonsmooth cost functions. *Advances in Neural Information Processing Systems*, 193–200.

Carmona CJ, Ramírez-Gallego S, Torres F, Bernal E, del Jesus MJ, García S (2012) Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems Appl.* 39(12):11243–11249.

Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.

Chae I, Bruno HA, Feinberg FM (2019) Wearout or weariness? Measuring potential negative consequences of online ad volume and placement on website visits. *J. Marketing Res.* 56(1):57–75.

Chakrabarti D, Kumar R, Radlinski F, Upfal E (2008) Mortal multi-armed bandits. *Adv. Neural Inform. Processing Systems* 21:273–280.

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 2249–2257.

Chen J, Yang X, Smith RE (2016) The effects of creativity on advertising wear-in and wear-out. *J. Acad. Marketing Sci.* 44(3):334–349.

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

Chen Y, Yang B, Dong J, Abraham A (2005) Time-series forecasting using flexible neural tree model. *Inform. Sci.* 174(3-4):219–235.

Chernev A, Hamilton R (2009) Assortment size and option attractiveness in consumer choice among retailers. *J. Marketing Res.* 46(3):410–420.

Feit EM, Berman R (2019) Test & roll: Profit-maximizing A/B tests. *Marketing Sci.* 38(6):1038–1058.

Foster D, Rakhlin A (2020) Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *Proc. Internat. Conf. on Machine Learn.*, 3199–3210.

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proc. Internat. Conf. on Machine Learn.*, 1050–1059.

Gilks WR, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. *J. Royal Statist. Soc. Ser. C* 41(2):337–348.

Goic M, Olivares M (2019) Omnichannel analytics. *Operations in an Omnichannel World* (Springer, Berlin), 115–150.

Goic M, Álvarez R, Montoya R (2018) The effect of house ads on multichannel sales. *J. Interactive Marketing* 42:32–45.

Goić M, Jerath K, Kalyanam K (2022) The roles of multiple channels in predicting website visits and purchases: Engagers versus closers. *Internat. J. Res. Marketing* 39:3.

Goic M, Rojas A, Saavedra I (2021) The effectiveness of triggered email marketing in addressing browse abandonments. *J. Interactive Marketing* 55:118–145.

Goldstein A, Hajaj C (2022) The hidden conversion funnel of mobile vs. desktop consumers. *Electronic Commerce Research and Applications*, 101135.

Han Y, Wang Y, Chen X (2021) Adversarial combinatorial bandits with general non-linear reward functions. *Proc. Internat. Conf. on Machine Learn.*, 4030–4039.

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proc. IEEE Internat. Conf. Comput. Vision* (IEEE, New York), 1026–1034.

Kahn BE, Wansink B (2004) The influence of assortment structure on perceived variety and consumption quantities. *J. Consumer Res.* 30(4):519–533.

Kireyev P, Pauwels K, Gupta S (2016) Do display ads influence search? Attribution and dynamics in online advertising. *Internat. J. Res. Marketing* 33(3):475–490.

Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *Proc. 44th Annual IEEE Sympos. on Foundations of Comput. Sci.* (IEEE, New York), 594–605.

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 8:30–37.

Koulouriotis DE, Xanthopoulos A (2008) Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Appl. Math. Comput.* 196(2):913–922.

Kuleshov V, Precup D (2014) Algorithms for multi-armed bandit problems. Preprint, submitted February 25, https://arxiv.org/abs/1402.6028.

Levine N, Crammer K, Mannor S (2017) Rotting bandits. *Adv. Neural Inform. Processing Systems*, 30.

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proc. 19th Internat. Conf. World Wide Web* (ACM, New York), 661–670.

Li L, Chu W, Langford J, Wang X (2011) Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proc. 4th ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 297–306.

Manchanda P, Dubé JP, Goh KY, Chintagunta PK (2006) The effect of banner advertising on internet purchasing. *J. Marketing Res.* 43(1):98–108.

Marszałkowski J, Drozdowski M (2013) Optimization of column width in website layout for advertisement fit. *Eur. J. Oper. Res.* 226(3):592–601.

Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing Atari with deep reinforcement learning. Preprint, submitted December 19, https://arxiv.org/abs/1312.5602.

Montgomery AL, Smith MD (2009) Prospects for personalization on the internet. *J. Interactive Marketing* 23(2):130–137.

Oh M, Iyengar G (2021) Multinomial logit contextual bandits: Provable optimality and practicality. *Proc. AAAI Conf. on Artificial Intelligence*, vol. 35, 9205–9213.

Ontanón S (2017) Combinatorial multi-armed bandits for real-time strategy games. *J. Artificial Intelligence Res.* 58:665–702.

Pandey S, Agarwal D, Chakrabarti D, Josifovski V (2007) Bandits for taxonomies: A model-based approach. *Proc. 2007 SIAM Internat. Conf. Data Mining* (SIAM, Philadelphia), 216–227.

Park CH, Park Y-H (2016) Investigating purchase conversion by uncovering online visit patterns. *Marketing Sci.* 35(6):894–914.

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, vol. 703 (John Wiley & Sons, Hoboken, NJ).

Riquelme C, Tucker G, Snoek J (2018) Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. Preprint, submitted February 26, https://arxiv.org/abs/1802.09127.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. (New Ser.)* 58(5):527–535.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4): 321–340.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536.

Russac Y, Vernade C, Cappé O (2019) Weighted linear bandits for non-stationary environments. *Adv. Neural Inform. Processing Systems*, 32.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.

Rutz OJ, Bucklin RE (2012) Does banner advertising affect browsing for brands? Clickstream choice model says yes, for some. *Quant. Marketing Econom.* 10(2):231–257.

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

Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural Networks* 61:85–117.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Sci.* 36(4):500–522.

Seznec J, Locatelli A, Carpentier A, Lazaric A, Valko M (2019) Rotting bandits are no harder than stochastic ones. *22nd Internat.*

Conf. Artificial Intelligence Statistics (PMLR, Long Beach, CA), 2564–2572.

Shepperd M, Cartwright M (2001) Predicting with sparse data. *IEEE Trans. Software Engrg.* 27(11):987–998.

Simchi-Levi D, Xu Y (2022) Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Math. Oper. Res.* Forthcoming.

Simonson I, Tversky A (1992) Choice in context: Tradeoff contrast and extremeness aversion. *J. Marketing Res.* 29(3):281–295.

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations Trends Machine Learn.* 12(1–2):1–286.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Machine Learn. Res.* 15(1):1929–1958.

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction, 2nd ed. (MIT Press, Cambridge, MA).

Swaminathan A, Joachims T (2015) Counterfactual risk minimization: Learning from logged bandit feedback. *Proc. Internat. Conf. on Machine Learn.*, 814–823.

Tang L, Rosales R, Singh A, Agarwal D (2013) Automatic ad format selection via contextual bandits. *Proc. 22nd ACM Internat. Conf. Inform. Knowledge Management* (ACM, New York), 1587–1594.

Tang L, Jiang Y, Li L, Zeng C, Li T (2015) Personalized recommendation via parameter-free contextual bandits. *Proc. 38th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 323–332.

Thomas P, Brunskill E (2016) Data-efficient off-policy policy evaluation for reinforcement learning. *Proc. Internat. Conf. on Machine Learn.*, 2139–2148.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Tversky A, Simonson I (1993) Context-dependent preferences. *Management Sci.* 39(10):1179–1189.

van Emden, Kaptein M (2018) Contextual: Evaluating contextual multi-armed bandit problems in R. Preprint, submitted July 8, https://arxiv.org/abs/1811.01926.

Verhoef PC, Kannan PK, Inman JJ (2015) From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *J. Retailing* 91(2):174–181.

Vermorel J, Mohri M (2005) Multi-armed bandit algorithms and empirical evaluation. *Proc. Eur. Internat. on Machine Learn.* (Springer, Berlin), 437–448.

Villar SS, Bowden J, Wason J (2015) Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statist. Sci.* 30(2):199.

Wang Y, Ouyang H, Wang C, Chen J, Asamov T, Chang Y (2017) Efficient ordered combinatorial semi-bandits for whole-page recommendation. *Proc. AAAI Conf. Artificial Intelligence*, vol. 31, no. 1, 2746–2753.

Wen Z, Kveton B, Ashkan A (2015) Efficient learning in large-scale combinatorial semi-bandits. *Proc. Internat. Conf. on Machine Learn.*, 1113–1122.