Sequential personalized menu optimization through bandit learning

Song, Xiang; Atasoy, B.; Ben-Akiva, Moshe E.

Publication date
2019

Document Version
Accepted author manuscript

Citation (APA)
Song, X., Atasoy, B., & Ben-Akiva, M. E. (2019). *Sequential personalized menu optimization through bandit learning*. Paper presented at 98th Annual Meeting of the Transportation Research Board (TRB), Washington D.C., District of Columbia, United States.

Sequential Personalized Menu Optimization through Bandit Learning

Xiang Song
MIT, Civil and Environmental Engineering
77 Massachusetts Avenue, Room: 1-249, Cambridge, 02139, MA
Tel: 857 200 7368; Email: bensong@mit.edu

Bilge Atasoy, Corresponding author
Delft University of Technology
Department of Maritime Transport and Technology
Mekelweg 2, 2628 CD, Delft, The Netherlands
b.atasoy@tudelft.nl

Moshe Ben-Akiva
Edmund K. Turner Professor of Civil and Environmental Engineering, MIT
77 Massachusetts Avenue, Room: 1-181, Cambridge, 02139, MA
Tel: 617 253 5324; Email: mba@mit.edu

Word count: 5611 words + 0 tables = 5611 words

Submission Date: November 15, 2018

ABSTRACT

This paper presents a sequential personalized menu optimization problem in the context of a Smart Mobility system that offers personalized menu of travel alternatives for each incoming traveler. The Smart Mobility system of interest is considered to combine together existing and emerging public and private transport alternatives. This paper extends existing literature on personalized menu optimization which was a static optimization problem to a sequential decision making under uncertainty problem. It unifies the preference learning and personalized menu optimization so that each time the traveler makes a choice, preference parameters are improved and the next menu optimization is done based on the updated preferences. In order to solve this problem, we propose a novel algorithm based on existing multi-armed bandit studies that address the trade-off between exploitation (offer optimal menu based on current belief) and exploration (experiment other menus if the current optimum is wrong). Numerical experiments show that our approach performs better than the classical heuristic. In addition, we compare it against static personalized menu optimization solution and find that exploration is needed under disturbance with inter-and intra-consumer heterogeneity.

Keywords: Smart Mobility, recommender systems, personalized menu optimization, sequential decision making under uncertainty, multi-armed bandit

INTRODUCTION AND MOTIVATION

Advances in information and communication technology (ICT) have been speeding up the emergence of innovative app-based transportation systems that provide different flexibilities. Uber, Lyft, and Zipcar are examples of such app-based services that distinguish themselves from traditional mobility systems with different characteristics. As they have addressed important travel needs they have been successful in attracting travelers *(1)*. These types of innovative services are also named under the concept of *Smart Mobility* as the operations are automatized and real-time data is used for real-time decisions *(2)*.

In order to design Smart Mobility systems, an innovative recommender system which can integrate both travel behavioral modeling and optimization techniques is often needed to achieve both personalization and efficiency. Such an innovative recommender system often offers individual traveler a personalized and optimized menu and we call such model as personalized menu optimization model. These models, though relatively new to transportation, have been developed and successfully applied in Smart Mobility systems such as Flexible Mobility on Demand (FMOD) and Tripod. Flexible Mobility on Demand (FMOD) is an app-based transportation service that is designed to provide personalized and optimized travel menus in real-time *(3)*. FMOD includes both private and public alternatives and is tested with simulation experiments and the presentation of optimized menus based on different objectives is shown to improve operator's profit and/or users' benefit *(4)*. Tripod is an app-based smart mobility system that incentivizes travelers based on energy savings in order to increase the utilization of more energy efficient options *(5)*. Travelers make trip requests on Tripod app, and the user level optimization generates personalized menus as a list of travel options including mode, departure time, route alternatives together with trip-making as well as driving style. Those alternatives are presented with energy usage and travel incentives in the form of tokens to incentivize user for green travel options. The travel menu on Tripod app is presented in FIGURE 1 where alternatives under different mode groups are presented on different tabs together with various information. This figure also serves as an example about what we mean by a travel menu in the context of a Smart Mobility system.

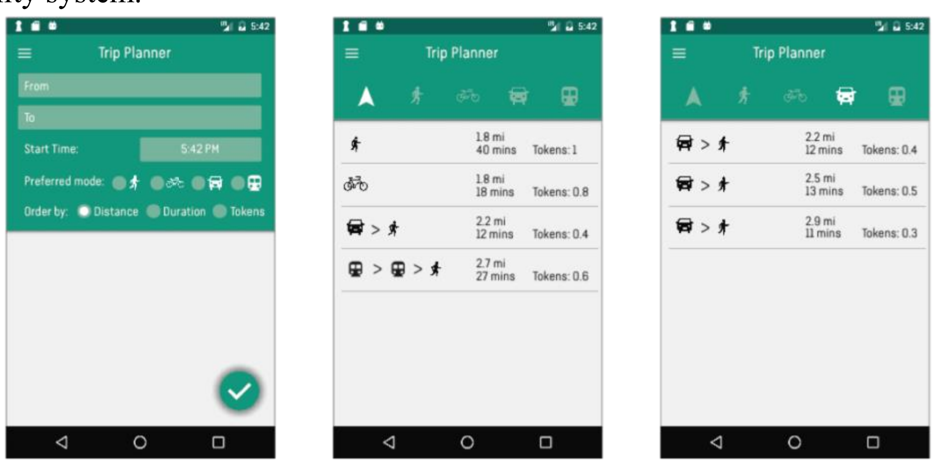

FIGURE 1 An example travel menu *(5)*

In order to enhance the performance of Smart Mobility using a personalized menu optimization model, advance behavioral models are crucial which can better capture real-life travelers' preferences by taking into account heterogeneity. A widely used type of choice models that consider inter-consumer heterogeneity is known as logit mixture (or mixed logit). Studies *(6,7)* have shown with simulation experiments and real data case study that the personalized menu optimization outperforms models that do not capture consumer heterogeneity well. Lately, a more advanced type of choice models, which captures both inter- and intra-consumer heterogeneity has been introduced and applied in economics, marketing and transportation *(8, 9,10)*.

In recent transportation related studies *(3, 7)*, personalized menu optimization (PMO) was studied where a customer's choice behavior is captured by a discrete choice model and the parameters of the choice model are inputs to the optimization model. However, in practice, the parameter value of choice model is often not known and has to be learnt gradually. Previously, we applied a preference updater *(11)* that is based on hierarchical Bayes (HB) estimator of logit mixture *(10)* to provide personalized menu optimization up-to-date estimates of. There, we didn't consider the learning of uncertain parameters when making the recommendation decision. However, there might be cases where current estimates from HB estimation procedure indicate car is the optimal alternative and PMO always offers car alternative but actually train is the favorite alternative of the consumer. In such a case, we need to go beyond *exploit-only* strategy, which is to offer an "optimal" menu based on current estimates of the parameters and *explore* other menus that may turn out to be optimal. Such problems that involve trade-off between *exploration* and *exploitation* is often formulated as a multi-armed bandit (MAB) problem.

In this paper, we propose a novel method called *UCB-Bayes* in order to learn the preference parameters of consumers while making travel decisions (on a smartphone app) in the context of a Smart Mobility system on a continuous basis. As the Smart Mobility system potentially includes emerging and innovative public and private alternatives, the learning of those parameters is critical to understand the response of travelers to those alternatives and provide them personalized travel menus. UCB-Bayes is built upon classical upper confidence bound (UCB) algorithm *(12)*. Particularly, we focus on the problem with menu size one in order to provide a proof-of-concept. The proposed method is novel with respect to existing MAB algorithm as its exploitation (or expected reward) is estimated by an HB estimator of logit mixture which differs from simple empirical mean in classical UCB algorithm *(12)*. Overall, we provide a unified framework for preference learning and personalized menu optimization that can be used in several Smart Mobility systems that are visited by travelers dynamically for travel recommendations.

The remainder of the paper is organized as follows. First, we review relevant literature and in the third section, we introduce the static and sequential personalized menu optimization problem along with logit mixture with inter-and intra-consumer heterogeneity. In the fourth section, we propose a novel solution algorithm along with benchmark solution algorithms. In the fifth section, numerical experiments are presented to illustrate the added value of the proposed method. We conclude our work and provide future directions in the last section.

LITERATURE REVIEW

In this section, we introduce relevant literature including consumer heterogeneity and choice models, assortment optimization and personalized menu optimization and multi-armed bandit problem.

Consumer Heterogeneity and Choice Models

Choice models that can well account for consumer heterogeneity is crucial for recommender systems. There exist various recommendation techniques that account explicitly for heterogeneity in consumer preferences *(13,14)*.

In the literature it is common to focus on the inter-consumer heterogeneity where the assumption is that consumers have stable individual preferences. However, consumer choices from repeated menus in laboratory and market experiments often deviates from neoclassical theory. There are a number of possible reasons including preferences may be situational, anchored or adapted to the status quo, and sensitive to context *(9)*.

Therefore, when we are doing dynamic demand forecasting that follows individuals over time, we need consumer models with both inter- and intra- consumer heterogeneity. There are a number of papers that introduce a structural system with both inter- and intra-consumer heterogeneity *(8, 9,10,12).*

Assortment Optimization and Personalized Menu Optimization

We have already seen a few papers in transportation, such as FMOD *(3,4),* where the recommendation problem can be formulated as an assortment optimization problem. Assortment optimization is an important problem in operations management and becomes popular in many practical settings such as retailing and online advertising *(16)*. In assortment optimization, different discrete choice models have been used to model the choice behavior of consumers including multinomial logit, nested logit, and logit mixture *(16,17,18)*. The goal of assortment optimization is to select a subset of items to offer from a universe of substitutable items in order to maximize the expected revenue when consumers exhibit a random choice behavior. We refer to Kök et al. *(19)* for more details of assortment optimization literature and industry practice.

Multi-armed Bandit Problem

Multi-armed bandit approach deals with the trade-off between exploitation (offer best alternatives based on current belief) and exploration (learning consumer's uncertain preferences of some alternatives) where recommendation decision is endogenous to preference updates. A typical MAB problem can be stated as follows *(20)*: there are N arms, each having an unknown success probability of emitting a unit reward. The success probabilities of the arms are assumed to be independent of each other. Many policies have been proposed under independent-arm assumptions *(21,12)*. Related with personalized menu optimization, the arm is the offered menu which is a list of alternatives and the success means an alternative being chosen by the consumer.

In this paper, we focus on the case where the menu size is one and therefore the arms are independent. If menu size is greater than one, the success probability of one arm/menu will depend on utility of multiple alternatives which means its reward is dependent on some of the other arms which has same alternatives on the menu. It is a combinatorial bandit problem where existing techniques such as UCB do not work directly on these functions *(22)*. We leave this more complicated case for future studies.

There are different types of MAB problems including stochastic, adversarial, and Markovian depending on the assumed nature of reward process *(23).* MAB problems usually do not have exact solutions except for some special cases *(24)* and many researchers have proposed different solution algorithms to different types of MAB problems: 1) *First explore then exploit* used by Rusmevichientong et al. *(25)* and Saure and Zeevi *(26)* to solve dynamic assortment optimization problems. 2) *Epsilon-greedy* with epsilon probability, choose a random arm to explore, otherwise exploit. 3) *Gittins index*, compute a Gittins index for each arm and choose the arm with highest index *(20)*. 4) *Randomized probability matching (RPM)*, randomly choose an arm with the probability that this arm is the best. A well-known special case of RPM is Thompson sampling (TS). 5) *Upper confidence bound (UCB),* choose an arm with the highest upper confidence bound. It has been applied in many fields including personalized recommendation in news articles *(27)* and digital coupon *(28)*.

Most existing literature in MAB field does not deal with discrete choice models but often assumes choice behavior follows simple Beta distribution *(28)*. In operations management, there exists literature proposing online policy depending on a priori knowledge of length of horizon *(25,26)* such as "first explore then exploit" policy. In MAB paradigm, Agrawal et al. *(29, 30)* propose an adapted TS method and a UCB method that can deal with multinomial logit choice model but relying on specific exploration phases.

The above-mentioned methods are not suitable for sequential personalized menu optimization setting where logit mixture is the underlying choice model. In this paper, we focus on proposing a method which adapts the classical UCB algorithm by utilizing the HB estimator for logit mixture of inter- and intra- consumer heterogeneity.

In transportation, there are a few studies about MAB problems which focus on different types of sequential decision-making problems. Chancelier et al. *(31)* have modeled route choice as a one-armed bandit problem (choice between a random and safe route) under different information regimes. They showed that risk neutral individuals tend to select risky routes while risk-averse individuals choose safe routes more frequently. Ramosa et al. *(32)* model the route choice problem as a multi-agent reinforcement learning scenario. They analyzed how travel information provided from a mobile navigation app would impact the agent route choice decision using epsilon-greedy strategy that minimizes difference between chosen route and best route.

MODEL AND SOLUTION

In this section, we first describe sequential personalized menu optimization problem, and then introduce its solution methods.

Sequential Personalized Menu Optimization

Assume T is the operational horizon. At each time period, there are N arriving consumers. The operator needs to decide which menu to offer (or in our case which alternative to offer) based on choice/menu history. After the operator offers the menu, the consumers need to decide whether to choose the alternative or opt out (reject the menu). After consumers make their choices, the operator needs to update the history particularly the estimates of choice model parameters.

Let P_{jnt} denote the choice probability of alternative j for consumer n at time t. x_{jnt} is a binary variable, which is equal to 1 if alternative j is offered to consumer n at time t, and 0 otherwise. At time period t, operator needs to decide which alternative to be offered, among NC many alternatives, that will maximize the total expected hit. Note that, in order to represent previous and future time periods with respect to the current time t, we use the index τ.

$$\max_{x_{jn\tau},\forall \mathrm{j},\tau} \sum_{\tau=t}^{T} \sum_{n=1}^{N} P_{jn\tau} x_{jn\tau} \tag{1}$$

subject to

$$\sum_{j=1}^{NC} x_{jn\tau} = 1, \forall n, \forall \tau \tag{2}$$

At time t, the operator actually just needs to decide on x_{jnt} based on all the choice history until time $t-1$. Additionally, the choice probabilities in the future are estimated based on history including time t. This problem does not have an exact solution.

We can also think of the objective function as an attempt to get as close as possible to the optimal alternatives for each individual (given by a clairvoyant who knows all the true parameter values). Particularly, we want to choose a solution method that minimizes the discrepancy between the optimal menus by the clairvoyant (given by j^*_{nt}) and menu offered by the solution ($\mathrm{j}_{\mathrm{nt}}^{\mathrm{solution}}$). In other words, we maximize the matching rate as:

$$\max_{solution} \sum_{n=1}^{N} \frac{1\{j^*_{nt} = \mathrm{j}_{\mathrm{nt}}^{\mathrm{solution}}\}}{N} \tag{3}$$

In this study, we assume that the choice behavior follows logit mixture. For logit mixture with inter- and intra-consumer heterogeneity, the choice probability of alternative *j* for consumer *n* at time *t* is as follows:

$$P_{jnt}(\eta_{nt}) = \frac{\exp\left(\mathrm{u}_{\mathrm{jnt}}(\eta_{\mathrm{nt}})\right)}{1 + \exp(\mathrm{u}_{\mathrm{jnt}}(\eta_{\mathrm{nt}}))} \tag{4}$$

where $\mathrm{u}_{\mathrm{jnt}}(\eta_{\mathrm{nt}})$ denotes the utility based on individual-and choice situation -specific parameter η_{nt}.

For logit mixture with inter- and intra-consumer heterogeneity, the posterior is given as follows:

$$K(\mu, \zeta_n\ \forall n, \eta_{mn}\ \forall mn, \Omega_w, \Omega_b | d_n \forall n)$$
$$\propto \prod_{n=1}^{N}\left[\prod_{m=1}^{M_n}\left[\prod_{j=1}^{J_{mn}}\left[P_j(\eta_{mn})^{d_{jmn}}\right] h(\eta_{mn}|\zeta_n, \Omega_w)\right] f(\zeta_n|\mu, \Omega_b)\right] k(\Omega_w)k(\mu)k(\Omega_b), \tag{5}$$

where η_{mn} represents a menu-specific parameter for menu *m* and consumer n, which follows a (normal) distribution with mean ζ_n and variance Ω_w represented by h. ζ_n represents individual-level parameters for a specific consumer *n*, which follows a (normal) distribution with mean μ and variance Ω_b represented by *f*. k denotes prior distributions for parameters. d_{jmn} indicates the chosen alternative as a binary term and d_n is the choice history vector for consumer n. See more details in Becker et al. *(10)*.

The estimation of η_{nt} can be done based on previous $t-1$ time periods of choice history through five-step HB procedure presented in Becker et al. *(10)*. Since each time period has its own posterior estimates, we use $\eta_{nt}^{t-1,s}$ denoting s^{th} draw of $(t-1)^{\text{th}}$ estimation, which will be used for personalized menu optimization at time period *t*. Note that we consider a total of *S* draws in order to represent the posterior estimates provided by the Bayesian procedure.

Solution Methods

Let $r_{jnt} = P_j(\eta_{nt})$ denote the expected reward or "revenue" for the operator. For clairvoyant who knows all the true parameter values η^*_{nt}, the optimal menu for consumer *n* at time *t* will be

$$j^*_{nt} = arg\max_j P_{jnt}(\eta^*_{nt}) \tag{6}$$

The operator has posterior estimates based on $t-1$ periods of choice history. The expected reward for menu *j* at time *t* for consumer *n* is then denoted by $\overline{r_{jnt}}(\eta_{nt}^{t-1})$ and given as follows:

$$\overline{r_{jnt}}(\eta_{nt}^{t-1}) = \frac{1}{S}\sum_{s=1}^{S}\frac{\exp\left(u_{jnt}(\eta_{nt}^{t-1,s})\right)}{1+\exp(u_{jnt}(\eta_{nt}^{t-1,s}))} \quad (7)$$

If we consider exploit-only, we offer the alternative based on current knowledge to obtain the maximum immediate revenue as given in equation (8). We refer to this as typical personalized menu optimization (PMO).

$$j_{nt}^{\mathrm{PMO}} = arg\max_{j}\overline{r_{jnt}}(\eta_{nt}^{t-1}) \quad (8)$$

However, since the parameter estimates include uncertainty, the offered menu may not be optimal. In addition, offering menu *j* will not give us information of alternative specific constants of other alternatives. We need to balance exploitation (offer the best menu based on current knowledge) and exploration (try other menus that may be optimal). Exploration will help us learn uncertain parameter values and will be beneficial for the objective of maximizing clicks across the whole operational horizon.

In order to balance the exploration and exploitation, we borrow the idea from one of the most widely used MAB heuristic, UCB. It uses the sum of empirical mean and a confidence bonus. The empirical mean based on choice history is as follows:

$$\overline{r_{jnt}} = \frac{1}{\sum_{\tau=1}^{t-1} x_{jn\tau}}\sum_{\tau=1}^{t-1} r_{x_{jn\tau}} \quad (9)$$

where we abuse the notation of *r* to also denote the realization of reward based on the menu decision $x_{jn\tau}$. Note that here the denominator needs to be at least one, i.e., alternative *j* is offered at least once before time *t*. In our experiments, we take care of it by an initial set of iterations where we offer each alternative once.

Our method uses not only the empirical mean, but also consider an additional term, which represents uncertainty about the alternative. We call this additional term the 'confidence bonus' term and therefore we offer a menu for consumer *n* at time *t* as follows:

$$\mathrm{j}_{\mathrm{nt}}^{\mathrm{UCB}} = \arg\max_{j}\left\{\overline{r_{jnt}} + \frac{1}{t-1}\sqrt{\frac{c\ log(t)}{\sum_{\tau=1}^{t-1} x_{jn\tau}}}\right\} \quad (10)$$

where the second term presents the "power" of exploration and constant *c* is a tuning parameter which controls the magnitude of exploration.

Given HB estimator for logit mixture, we replace $\overline{r_{jnt}}$ by the estimated expected reward $\overline{r_{jnt}}(\eta_{nt}^{t-1})$ and call the algorithm *UCB-Bayes*, which chooses the menu as follows:

$$\mathrm{j}_{\mathrm{nt}}^{\mathrm{UCB-Bayes}} = \arg\max_{j}\left\{\overline{r_{jnt}}(\eta_{nt}^{t-1}) + \frac{1}{t-1}\sqrt{\frac{c\ log(t)}{\sum_{\tau=1}^{t-1} x_{jn\tau}}}\right\} \quad (11)$$

NUMERICAL EXPERIMENTS

Experimental Setup

In this section, we present numerical experiments under different conditions to evaluate the performance of different solution methods: PMO, UCB, and UCB-Bayes. We use 5 alternatives, and the utility of alternative *j* for consumer *n* at time *t* is given as:

$$\mathrm{u_{jnt}}(\eta_{\mathrm{nt}}) = \left(\alpha_{\mathrm{jnt}} - \exp\left(\beta_{\mathrm{tt,n,t}}\right)\mathrm{TT_{j,n,t}} - \mathrm{TC_{j,n,t}}\right)/\exp\left(\beta_{\mathrm{tc,n,t}}\right) \quad (12)$$

where $\eta_{nt} = (\alpha_{1nt}, \dots, \alpha_{Jnt}, \beta_{tt,n,t}, \beta_{tc,n,t})$ denotes the menu-specific parameter vector for user n. Index *t* denotes menu as at each time period one menu is offered. $(\alpha_{1nt}, \dots, \alpha_{Jnt})$ is the vector of alternative specific constants. $\beta_{tt,n,t}$ is the travel time coefficient and $\beta_{tc,n,t}$ is the travel cost coefficient which are both lognormally distributed. Alternative 5 is considered to be the base and therefore $\alpha_{5nt} = 0\ \forall n, t$. Utility is given in monetary value (willingness to pay space).

In the first five periods, we display alternative *t* for all the individuals (i.e., they see each alternative once) to warm up the system and obtain basic knowledge about alternatives. We construct a synthetic sample by drawing *N* times from the multivariate normal distribution associated with the individual-level parameters. For logit mixture with inter- and intra-consumer heterogeneity, we further draw the menu-specific parameters with individual-specific mean and covariance matrix for intra-consumer heterogeneity. At each time period, we offer one alternative for each consumer for different solution methods and compare whether the offered menu is the same as the optimal menu. Travel time and cost are drawn from Uniform [0,1] for every alternative *j*, consumer *n*, and time *t*. Tuning parameter, *c*, is set to 2 unless otherwise noted.

Experimental Results

Experiments comparing UCB-Bayes and UCB

In this section, we first compare UCB-Bayes and UCB methods. Two different sample mean vectors including (0, 0.5, 1, 1.5, -1, -1) and (0, 0.5, 0.8, 1, -1, -1) are used. Remind that the first 4 correspond to alternative specific constants of the first 4 alternatives and the last 2 parameters are time and cost coefficients, respectively. The covariances for inter- and intra- consumer heterogeneity are both represented by a diagonal matrix.

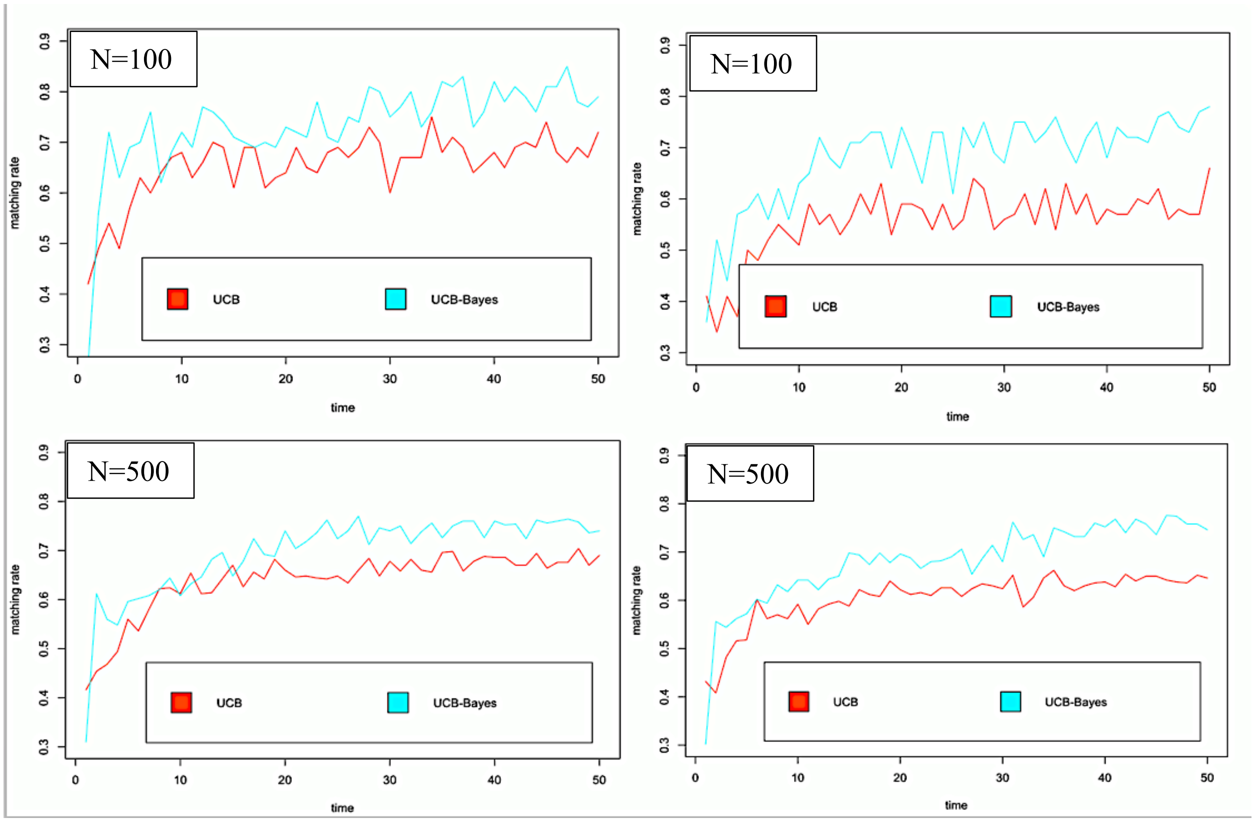

FIGURE 2 UCB versus UCB-Bayes under logit mixture with inter-consumer heterogeneity

In FIGURE 2 we compare UCB and UCB-Bayes where the y-axis denotes the matching rate (proportion of the cases where offered menus correspond to optimal menus) and x-axis denotes the time periods. Here, we consider logit mixture with inter-consumer heterogeneity only. The left column is associated with the set of parameters (0, 0.5, 1, 1.5, -1, -1) and right is with (0, 0.5, 0.8, 1, -1, -1). The upper ones are obtained using N=100 and the bottom ones are with N=500. We observe that both algorithms learn what are the optimal menus. The performance of UCB-Bayes is in general better under different conditions with a gap of around 10%.

Furthermore, we analyze logit mixture with inter- and intra-consumer heterogeneity, which means for a given individual, taste preferences vary across time periods, i.e., across different choice situations. It leads to a more difficult problem of learning the preferences. In FIGURE 3, we observe that UCB-Bayes outperforms UCB in general under different true sample mean vectors and sample sizes. However, the gap between the two methods in terms of matching rate is smaller than those under inter-consumer heterogeneity only.

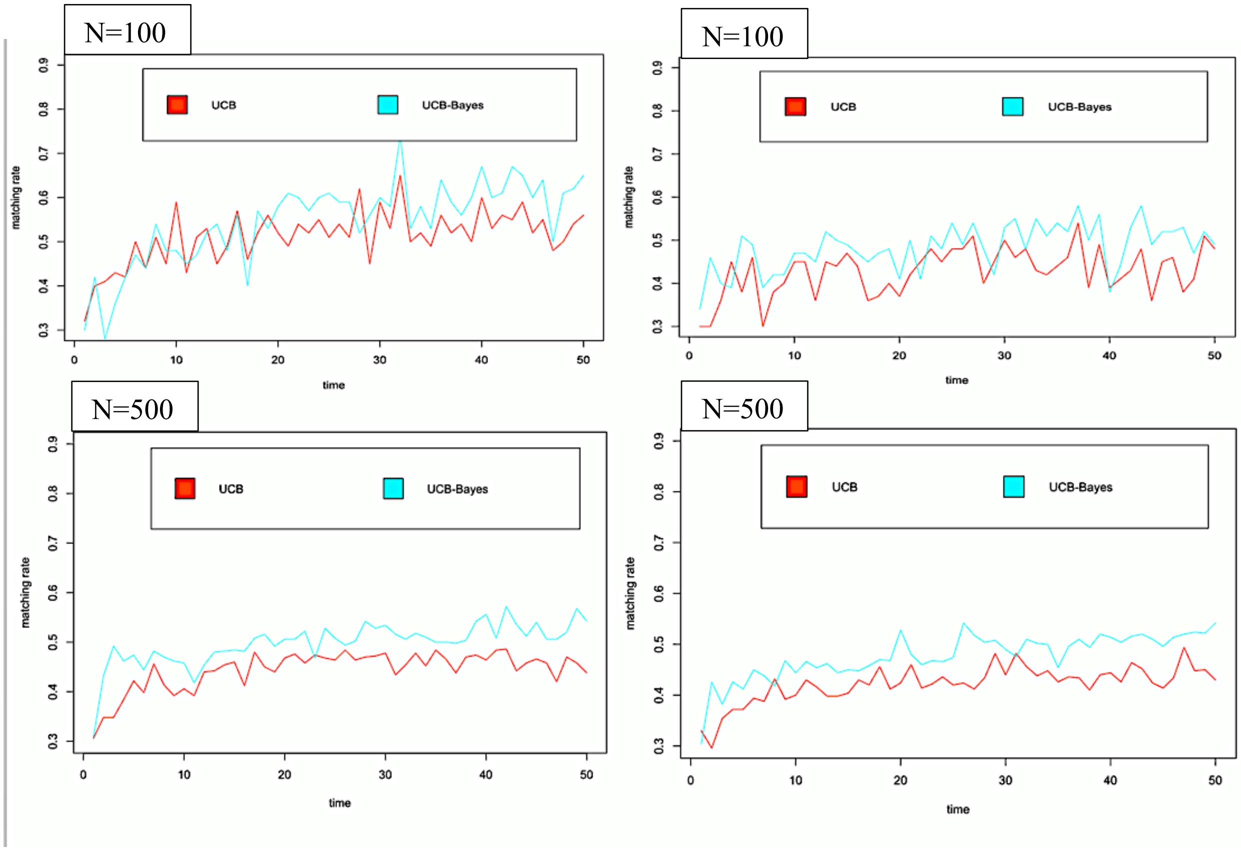

FIGURE 3 UCB versus UCB-Bayes under logit mixture with inter- and intra-consumer heterogeneity

Experiments comparing UCB-Bayes and PMO

In this section, we compare UCB-Bayes and PMO in order to evaluate the benefit of adding a confidence bonus term to explore beyond expected value predicted by the HB estimator.

Figure 4 illustrates the comparison between UCB-Bayes and PMO under logit mixture. The sample mean vector used is (1, 3, 5, 7, 1, -1). The top two shows cases where the variance is equal to the identity matrix (I). The bottom two show cases where variance is large (100 I). The left two show cases where the true choice model is logit mixture with inter-consumer heterogeneity. The right two show cases where the true choice model is logit mixture with inter-and intra-consumer heterogeneity.

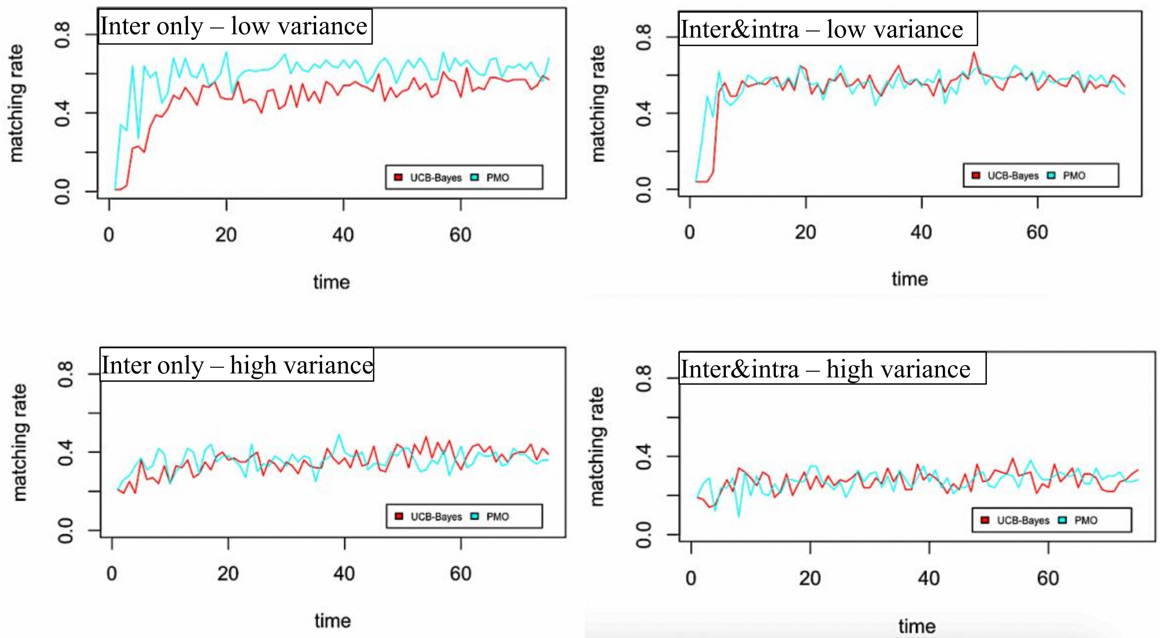

FIGURE 4 Comparison between UCB-Bayes and PMO

In FIGURE 4, we observe that PMO is better than UCB-Bayes when the true choice behavior has inter-consumer heterogeneity only, which means nullifying the confidence bonus (i.e., setting c=0) would be the best case for UCB-Bayes. One reason may be that PMO has collected enough information about each alternative and there is no need to explore beyond estimated best alternatives. UCB-Bayes' exploration makes it deviate more from the clairvoyant. When variance becomes large, the performance of both methods gets worse. With inter- and intra-consumer heterogeneity, the performances of the two methods are similar.

There might be cases where the optimal alternative is under disturbance for a certain period of time so that its attributes, e.g., travel time and travel cost, may be much worse than other alternatives. An exploit-only strategy, like PMO, might get trapped within suboptimal alternatives. In order to evaluate the benefits of exploration, we propose an alternative setting where optimal alternative is under disturbance and an exploit-only strategy would not be able to offer it. Particularly, in the first BT time periods, we draw the travel time and travel cost of alternative 4 (which is the most preferred alternative on average according to sample-level alternative specific constants) to be from Uniform [5,10] whereas they are drawn from Uniform [0,1] for other alternatives. Then, as of time period BT+1, we start to draw time/cost from Uniform [0,1] as other alternatives. FIGURE 5 illustrates the comparison under disturbance with logit mixture with inter-consumer heterogeneity. The left figure shows a case where the true variance matrix is assumed to be 0.1 I and for the right figure it is assumed to be I.

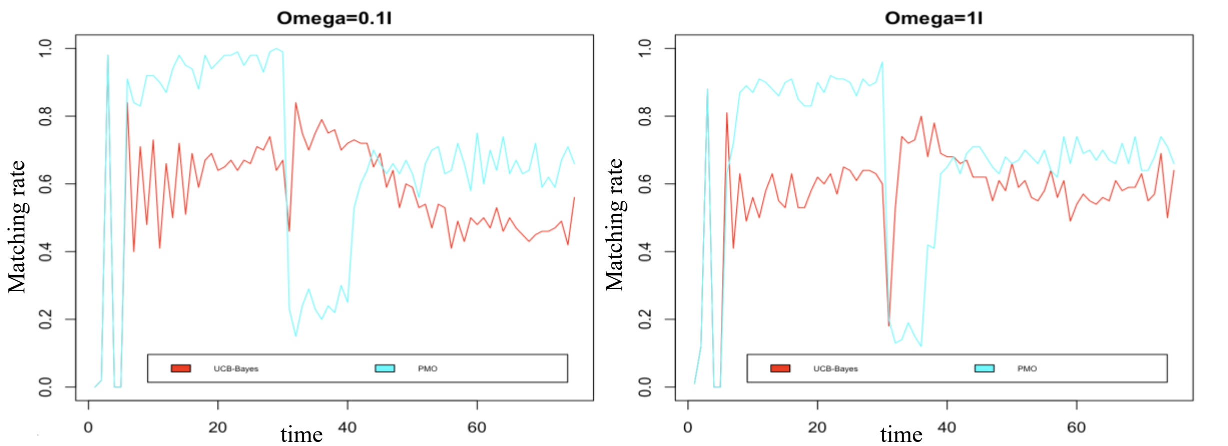

FIGURE 5 UCB-Bayes versus PMO under disturbance using logit mixture with inter-consumer heterogeneity, BT=30

During the disturbance, both methods rarely choose alternative 4. When the disturbance is over, both methods have big drops in their matching rates. For UCB-Bayes, the drop is quickly recovered and it performs better than PMO for several periods. It takes more time periods for PMO to recover and eventually both methods reach similar matching rates though PMO performs slightly better. The recovery is easier for PMO when variance is larger.

Furthermore, we consider cases where the true underlying choice model is logit mixture with inter- and intra-consumer heterogeneity. FIGURE 6 presents the comparison under disturbance with inter- and intra- consumer heterogeneity. The left and right four plots show cases where true variance is 0.1I and I, respectively. The four rows use different values of *c* as 0.5, 2, 5, and 10.

Different than cases with only inter-consumer heterogeneity, PMO may get trapped with suboptimal alternatives when there is also intra-consumer heterogeneity and therefore UCB-Bayes performs better. The performance gap between PMO and UCB-Bayes also depends on the level of variance, i.e., lower variance has negative impact on the performance of PMO.

When *c*=0, UCB-Bayes reduces to PMO. The magnitude of *c* controls how much we want to explore beyond PMO results. Large values of *c* may explore too much and result with bad menus. Therefore, under disturbance with inter- and intra-consumer heterogeneity, there would be an optimal value of *c*. In FIGURE 6, we observe that under different variances, different values of *c* perform the best. Under variance of 0.1I, both *c*=2 and *c*=5 perform better than larger or smaller values of *c*. Similarly, under variance of I, *c*=2 performs the best. In real life, the optimal tuning parameter can be found through splitting user traffic and experimenting different values of *c* to determine the degree of exploration.

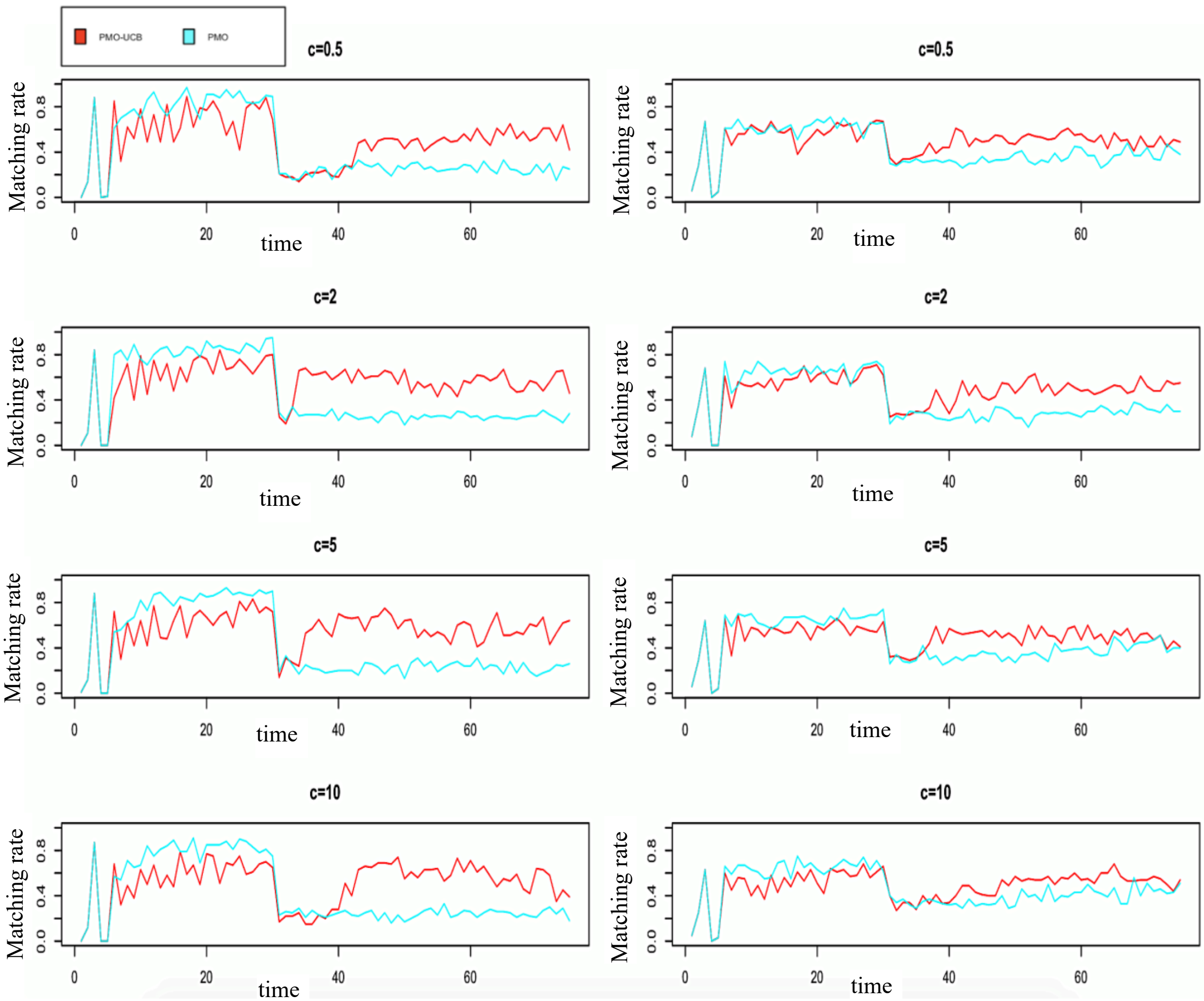

FIGURE 6 Comparison between UCB-Bayes and PMO under disturbance with logit mixture with inter-and intra-consumer heterogeneity, BT=30

CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel method, UCB-Bayes, for unified preference learning and personalized menu optimization in the context of Smart Mobility. UCB-Bayes adapts the classical UCB algorithm by using the HB estimates for logit mixture. The proposed algorithm outperforms the classical algorithm under various conditions. The performance gap becomes smaller when the true choice model is logit mixture with inter- and intra-consumer heterogeneity. We also compare the proposed algorithm with PMO and find that in regular settings, UCB-Bayes performs worse than PMO given that true choice model is logit mixture with only inter-consumer heterogeneity. In other words, in such settings UCB-Bayes reduces to PMO without any exploration term. This happens as it explores when the estimates are already good. On the other hand, when intra-consumer heterogeneity is also considered, the performance of the two methods becomes similar.

Under an alternative setting where there is disturbance for a certain time frame, which prohibits system operators to offer optimal alternatives (e.g., closure of a road, subway system etc.), the performance of PMO is negatively affected. Especially when the true underlying model is logit mixture with inter-and intra-consumer heterogeneity, PMO performs worse than UCB-Bayes. This indicates that more exploration is needed under disturbance. The magnitude of heterogeneity also has an impact on the relative performance of the two methods.

In summary, when we believe the consumer heterogeneity among consumers is not high and intra-consumer heterogeneity exists, we propose to use UCB-Bayes especially when there exists some disturbance for some alternatives. In other cases, PMO might perform better, i.e., exploration may not be needed.

In the future, we need to investigate realistic cases where menu size is greater than one and therefore the rewards of different menus are correlated. It requires a different algorithm and its combinatorial nature would make it computationally difficult to choose among many possible menus. Furthermore, the application of the proposed framework in real case studies is a very interesting direction to take as the heterogeneity will be coming from real choices of individuals across the population and the framework can be validated.

ACKNOWLEDGMENT

This work is in the context of a research project, called Sustainable Travel Incentives with Prediction, Optimization and Personalization (Tripod), sponsored by the U.S. Energy Department's Advanced Research Projects Agency-Energy (ARPA-E). It was awarded by ARPA-E's one of the latest programs, the Traveler Response Architecture using Novel Signaling for Network Efficiency in Transportation (TRANSNET). We thank the reviewers for the comments. As the paper is not accepted for publication we did not address all of them for the TRB Annual Meeting 2019, but we addressed most of them. The extended abstract we submitted considers all the comments though.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: X. Song, B. Atasoy, M. Ben-Akiva; data generation: X. Song, B. Atasoy; analysis and interpretation of results: X. Song, B. Atasoy; draft manuscript preparation: X. Song, B. Atasoy. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Rayle, L., Shaheen, S., Chan, N., Dai, D., and Cervero, R. (2015). "App-based, on-demand ride services: Comparing taxi and ridesourcing trips and user characteristics in San Francisco". TRB Annual Meeting Compendium of papers.
2. Földes, D. and Csiszár, "Conception of Future Integrated Smart Mobility", *2016 Smart Cities Symposium Prague (SCSP),* Prague, 2016, pp. 1-6.
3. Atasoy, B., Ikeda, T., Song, X. and Ben-Akiva, M. (2015), "The Concept and Impact Analysis of a Flexible Mobility on Demand System", *Transportation Research Part C: Emerging Technologies*, Vol. 56, pp. 373-392.
4. Atasoy, B., Ikeda, T. and Ben-Akiva, M. (2015), "Optimizing a Flexible Mobility on Demand System", *Transportation Research Record (TRR),* Vol. 2536, pp. 76-85.
5. Azevedo, C. L., Seshadri, R., Gao, S., Atasoy, B., Akkinepally, A. P., Trancik, J. and Ben-Akiva, M. E. (2018) Tripod: Sustainable Travel Incentives with Prediction, Optimization and Personalization, TRB 97th Annual Meeting, Washington D.C., USA.
6. Song, X., Atasoy, B., and Ben-Akiva, M. (2017). Smart Mobility Through Personalized Menu Optimization. In Proceedings of the 96th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.
7. Song, X., Danaf, M., Atasoy, B., and Ben-Akiva, M. (2018). Personalized Menu

Optimization with Preference Updater: A Boston Case Study. Forthcoming at Transportation Research Record.
8. Hess, S., & Train, K. E. (2011). Recovery of inter-and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological*, 45(7), 973-990.
9. Ben-Akiva, M., McFadden, D., and Train, K. (2016). Foundations of stated preference elicitation, consumer choice behavior and choice-based conjoint analysis. Working paper, Department of Economics, University of California, Berkeley.
10. Becker, F., Danaf, M., Song, X., Atasoy, B., and Ben-Akiva, M.E. (2018). Bayesian estimator of logit mixture models with inter- and intra-consumer heterogeneity. Forthcoming at Transportation Research Part B: Methodological.
11. Danaf, M., Becker, F., Song, X., Atasoy, B., and Ben-Akiva, M.E. (2018). Personalized recommendations using discrete choice models with inter-and intra-consumer heterogeneity. Working paper, MIT.
12. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47, 235–256.
13. Allenby, G. and Ginter, J. (1995). Using extremes to design products and segment markets. *Journal of Marketing Research*, 32, 392–403.
14. Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, Vol. XXXVII, 363–375.
15. Désir, A., Goyal, V., Segev, D., & Ye, C. (2015). Capacity constrained assortment optimization under the Markov chain based choice model. *Operations Research*, Forthcoming.
16. Davis, J., Gallego, G., & Topaloglu, H. (2013). Assortment planning under the multinomial logit model with totally unimodular constraint structures. Department of IEOR, Columbia University. Available at http://www. columbia. edu/∼ gmg2/logit_const. pdf.
17. Davis, J. M., Gallego, G., & Topaloglu, H. (2014). Assortment optimization under variants of the nested logit model. *Operations Research*, 62(2), 250-273.
18. Feldman, J., & Topaloglu, H. (2015). Bounding optimal expected revenues for assortment optimization under mixtures of multinomial logits. *Production and Operations Management*, 24(10), 1598-1620.
19. Kök, A. G., Fisher, M. L., & Vaidyanathan, R. (2008). Assortment planning: Review of literature and industry practice. In Retail supply chain management (pp. 99-153). Springer US.
20. Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society, Series B: Methodological 41, 148-177.
21. Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6, 4–22.
22. Chen, W., Hu, W., Li, F., Li, J., Liu, Y., & Lu, P. (2016). Combinatorial multi-armed bandit with general reward functions. In Advances in Neural Information Processing Systems (pp. 1659-1667).
23. Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning, 5(1), 1-122.
24. Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. Applied Stochastic Models in Business and Industry 26, 639-658.

25. Rusmevichientong, P., Shen, Z. J. M., & Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. Operations research, 58(6), 1666-1680.
26. Sauré, D., & Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. Manufacturing & Service Operations Management, 15(3), 387-404.
27. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web (pp. 661-670). ACM.
28. Song, X. (2016). A Bayesian bandit approach to personalized online coupon recommendations MSc Thesis, Massachusetts Institute of Technology.
29. Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2017a). MNL-Bandit: A Dynamic Learning Approach to Assortment Selection. arXiv preprint arXiv:1706.03880.
30. Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2017b). Thompson Sampling for the MNL-Bandit. arXiv preprint arXiv:1706.00977.
31. Chancelier, J. P., De Lara, M., & De Palma, A. (2007). Risk aversion, road choice, and the one-armed bandit problem. Transportation Science, 41(1), 1-14.
32. Ramosa, G. D. O., Bazzana, A. L., & da Silvaa, B. C. (2018). Analysing the impact of travel information for minimising the regret of route choice. Transportation Research Part C: Emerging Technologies, 88, 257-271.