

**Core Methodological Development in**  
**“Bandit’s Paradise: Customer Acquisition through Online Display Advertising”**

Eric M. Schwartz \*

October 21, 2012

**Purpose of document**

This document is a working draft combining content from two essays of my dissertation: one that focuses on a large-scale field experiment in which I allocated online display advertisements across websites using state-of-the-art multi-armed bandit (MAB) algorithms; the other focuses on the methodological development and the general background of the core approach used for those allocations.

The methods described here cover, but are not strictly limited to, the solution procedure used in my field experiment. I characterize the features of the real-world advertisers’ problems that motivate this work (and are present in many other kinds of dynamic allocation problems). These features require a novel version of the MAB framework: the hierarchical attribute-based batched MAB. Choosing among existing families of MAB solutions, I employ randomized probability matching, which is better suited than other solution algorithms to accommodate a hierarchical Bayes choice model (which is used in the experiment).

The empirical results from the experiment are not discussed in detail here, but they will be highlighted substantially in my dissertation essays.

I will send updated versions of the drafts as they are finished.

## **Bandit's Paradise: Customer Acquisition through Online Display Advertising**

Eric M. Schwartz \*

Eric T. Bradlow

Peter S. Fader

October 21, 2012

### **Abstract**

As business experiments such as A/B or multivariate tests become more popular, firms seek to move beyond repeated “testing and learning” towards “earning while learning.” For instance, online advertisers deliver dozens of different ad executions across dozens of websites with the goal of acquiring customers. Over time, they want to adapt and allocate relatively more impressions to the better performing ads on each website, but they want to be profitable even while they are testing. Since the the experiment is meant to be profitable, we frame this as solving the multi-armed bandit problem. The contribution of the paper is to extend existing methods for bandit problems to handle various components of real-world marketing experiments (e.g., attributes of actions, unobserved differences in context, batched decisions). The key methodological innovation is applying hierarchical/multilevel models to randomized probability matching, allowing us to solve the batched hierarchical attribute-based multi-armed bandit problem. The benefits of this new approach are demonstrated through a field experiment with a large retail bank using online advertising to acquire customers. The approach is also compared to benchmark methods via simulation studies.

\* Eric M. Schwartz is a doctoral candidate in Marketing at the Wharton School at the University of Pennsylvania. This is part of the first author's dissertation, which is co-advised by Eric T. Bradlow, the K. P. Chao Professor, Professor of Marketing, Statistics and Education, Vice-Dean and Director of Wharton Doctoral Programs, and Co-Director of the Wharton Customer Analytics Initiative; and Peter S. Fader, the Frances and Pei-Yuan Chia Professor, Professor of Marketing, and Co-Director of the Wharton Customer Analytics Initiative. The author thanks seminar and conference participants at Erasmus University, Marketing Science 2012, Marketing in Israel XI 2011, Tilburg University, and the University of Pennsylvania. The authors thank the Wharton Customer Analytics Initiative and the Wharton Risk Management and Decision Processes Center for its support through the Russell Ackoff Doctoral Student Fellowships. All correspondence on this manuscript should be addressed to Eric M. Schwartz, ericschw@wharton.upenn.edu, 215-573-0539; 3730 Walnut Street, Jon M. Huntsman Hall, Suite 700, Philadelphia, PA, 19104.

# 1 Introduction and motivation

As business experiments such as A/B/C or multivariate tests become more popular, firms aim to continuously be “testing and learning” about their market environments. But as this practice becomes part of regular business operations, such continuous testing has to be done profitably – i.e., to be “earning while learning.” For instance, many online advertisers regularly deliver dozens of different display ad executions across dozens of websites in order to achieve measurable business goals, e.g., customer acquisition. Over time, they aim to adapt and allocate more impressions to the better performing ads on each website.

This problem is not at all unique to advertisers; it belongs to a much broader class of problems that decision-makers face across countless domains. In marketing alone, many other activities – sending emails, customer service policies, designing websites, recommending products – can be framed as sequential and adaptive experiments. Such domains make up a rich class of direct/interactive marketing problems structured around the question: which targeted marketing action should we take, when, with which customers, and in which contexts?

Profitable experimentation illustrates the dilemma of exploration versus exploitation, where a manager chooses between an action that is thought to be performing well currently (*exploit*) and an uncertain action with the hope of benefiting from that learning in the future (*explore*). Since the the experiment is meant to be profitable, we frame this as solving *multi-armed bandit problem* (MAB). Solving this dilemma between exploration and exploitation (*learning and earning*) has been the focus of much research.

The MAB is a sequential decision-making problem with the goal of gaining the largest possible cumulative reward, which comes from a set of payoff distributions, each with unknown parameters. Its name references the problem of a gambler who wants to maximize his returns from repeatedly playing several slot machines (aka “one-armed bandits”) but lacking precise initial knowledge of each machine’s payoff rate. We refer to this as the *basic* MAB. At each step, we decide which arm to play next and observe only the reward from that arm. Then we adapt by updating our beliefs and select which arm we would like to play next. This problem has a long history from statistics (Berry 1972; Bradt et al. 1956; Robbins 1952; Thompson 1933), economics (Rothschild 1974), and operations research (Gittins 1979; Gittins and Jones

1974). It even serves as the one of the fundamental problems in the field of computer science, i.e., *reinforcement learning*, which is concerned with agents earning rewards by interacting with their environments (Sutton and Barto 1998). Advances in all these fields have developed methods to solve more sophisticated versions of the bandit problem by adding components to extend the basic MAB, as we will discuss in detail. However, no extension has fully embraced all of the features of the large-scale MAB that real-world decision-makers (such as online advertisers) face on a routine basis. These features include

1. *Attributes*. Actions are characterized by a known attribute structure, so they cannot be treated as if they were independent of each other, i.e., payoffs of different actions may be interdependent. This attribute-based setting is often known as a linear bandit.
2. *Unobserved heterogeneity* across contexts. There is a hierarchical structure in which the MAB is replicated in each context, but the response to the same action may differ across contexts in unobserved ways.
3. *Batched decisions*. At each decision period, many observations must be allocated across actions, requiring allocation probabilities instead of a one-at-a-time selection of actions. Batching is sometimes known as delayed feedback.

These three features correspond to how we described the real-world adaptive advertising allocation problem in the introduction, and are common in many other large-scale decision problems. Therefore, we want to solve the explore/exploit dilemma in the presence of the above three challenges; unfortunately, previous methods only allow us to solve a much more limited set of problems. While various papers (to be reviewed shortly in Table 1) have addressed each of these features separately and little work deals with any two of them, we are unaware of any work that addresses all three in a comprehensive but practical manner.

The contribution of the paper is to extend existing methods to handle various components of real-world marketing experiments (e.g., attributes of actions, unobserved differences across contexts, batched decisions). Hence, we broaden the class of bandit problems that we can solve well in practice. The key methodological innovation is applying hierarchical/multilevel models to *randomized probability matching* as a solution to the *hierarchical attribute-based*

*multi-armed bandit problem with batching*. This approach accumulates more reward (e.g., more customers acquired) and identifies the best action in each context faster (e.g., best ad per website) than existing methods when applied to this problem.

The rest of this document is structured around versions of the MAB, each with a different combination of the three key features discussed, and the solutions appropriate for each of those problems. We begin by formalizing the MAB that we want to solve as a classic optimization problem, i.e., as a Markov decision process. However, we note why it cannot be solved directly with existing dynamic programming methods, so we strip it down to the *basic* MAB, which is the only version examined that does have an exactly optimal solution, and is thus very narrowly defined. In addition, various approximate solutions are discussed. One such family of solutions that is more generalizable to other MAB settings is known as *upper confidence bound* policies. We then rebuild the problem of interest feature-by-feature and formalize the appropriate solutions for the versions of the MAB along the way. Next comes the *attribute-based* MAB, which does have an asymptotically optimal solution based on extensions of the approximations to the basic MAB. Finally, by adding in a layer of unobserved heterogeneity across contexts (i.e., hierarchical structure), as well as batched decision making, we arrive at the problem we want to solve, the *batched hierarchical attribute-based bandit*. We conclude by discussing why the available approximation methods are not applicable here, and we employ another solution approach based on the principle of *randomized probability matching*.

## 2 Problem formulation

### 2.1 Actions, contexts, rewards, and adaptive allocation

Suppose we can employ  $K$  actions indexed by  $k = 1, \dots, K$  in each of  $J$  contexts indexed by  $j = 1, \dots, J$ . Each action is characterized by  $d$  attributes, denoted by vector  $x_k = (x_{k1}, \dots, x_{kd})$ , known and observed a priori for all actions. At the start of each period  $t$ , within each context  $j$ , we decide how to allocate a batch of  $M_{jt}$  observations across the  $K$  actions. That is, we design an experiment where each action  $k$  will be observed  $m_{jkt}$  times, so we describe the proportion of observations allocated to that action by action weights,  $w_{jkt} = m_{jkt}/M_{jt}$ , with the batch size

constraint  $1 = \sum_{k=1}^K w_{jkt}$ .

The action weights  $w_{jkt}$  for all  $j$  and  $k$  are instructions for how to allocate the observations in period  $t$ . After those weights are set, we observe rewards for all context-action combinations at the end of the period, and we then adapt and adjust the weights for the next period  $t + 1$ . Therefore, this is viewed as the setup for the sequential allocation problem (also known as a sequential sampling problem or an adaptive design of experiments problem).

The reward of taking action  $k$  in context  $j$  during period  $t$  is a realization of the random variable,  $Y_{jkt}$ . We assume the rewards  $\{y_{jk1}, y_{jk2}, \dots, y_{jkt}\}$  are independently and identically distributed following  $f(y|x_k, \beta_j)$ .

$$\{y_{jk1}, \dots, y_{jkt}\} \sim f(y|x_k, \beta_j) \forall j, k \quad (1)$$

The expected rewards,  $E_f(y|x_k, \beta_j)$ , differ due to both the action’s attributes  $x_k$  and the context-specific parameter vector  $\beta_j$ . The attribute-based differences are described by a common set of  $d$  attributes, so rewards arising from different actions can be correlated. This feature of a MAB is often known as a “linear bandit” (Dani et al. 2008; Rusmevichientong and Tsitsiklis 2010), but we refer to it more generally as an “attribute-based bandit.” This then yields,

$$\mu_{jk} = E_f(y|x_k, \beta_j) = \text{link}^{-1}(x'_k \beta_j) \forall j, k, \quad (2)$$

where  $\text{link}^{-1}()$  is inverse of any link function common in generalized linear models (e.g., log, log-odds).

The expected reward for action  $k$  may differ across the  $J$  contexts. Even though the action attribute vector,  $x_k$ , is observed, we let  $\beta_j$  be the unobserved parameter vector that is context-specific, denoting the impact of the  $d$  attributes in context  $j$ . We parameterize how  $\beta_j$  differs across contexts by positing a population-level distribution  $g(\bar{\beta}, \Sigma)$ . Therefore, the attribute-based bandit problems may have different solutions in each context (e.g., different rank ordering of the actions’ expected rewards). This component of the problem creates a hierarchical or multilevel structure. The same  $K$  actions are replicated across  $J$  context-specific

experiments, but those contexts are all considered to be draws from a larger population.

$$\beta_j \sim g(\bar{\beta}, \Sigma) \forall j \quad (3)$$

All of the parameters in this problem are described by  $\theta = (\{\beta_j\}_{j=1}^J, \bar{\beta}, \Sigma)$ . While those are all of the parameters, we will pay special attention to the expected rewards,  $\mu_{jk}$ , and our current state of knowledge or “beliefs” about those expected rewards. However, any uncertainty that we have about the expected rewards is driven by our uncertainty in the components of  $\theta$ . We note that the  $\theta$ ,  $x_k$ , and  $f(\cdot)$  are assumed to be stationary over time for each context-action combination. Relaxing the assumption of stationarity would lead to the “restless bandit” (Whittle 1980). This is the MAB where unknown expected rewards are not stationary but the three features discussed are not present. However, we do not pursue it here.

The reward described thus far is from a single observation of action  $k$  in context  $j$ , but we are interested in describing the reward from a whole batch of  $M_{jt}$  observations allocated following weights  $w_{jt}$ . We denote this as simply as  $R(w_{jt}, \beta_j) = \sum_{k=1}^K Y_{jkt}$ . Therefore expected reward of a batch,

$$\mathbb{E}_f R(w_{jt}, \beta_j) = \sum_{k=1}^K w_{jkt} M_{jt} \mu_{jk} = \sum_{k=1}^K w_{jkt} M_{jt} \text{link}^{-1}(x'_k \beta_j) \forall j. \quad (4)$$

Later we will also discuss how well different methods are constructed to minimize cumulative *regret*. Regret is defined as the expected sum of how much better you could have played over time: the sum of differences between the optimal action’s expected reward and the expected reward of the actions chosen (hence, linear loss function) (Lai and Robbins 1985). Formally, for any action context  $j$ , there exists an action with expected reward  $\mu_{j*} = \max\{\mu_{j1}, \dots, \mu_{jK}\}$ , known as the optimal level of reward. Then regret of any action is defined as the expectation of the difference between the best possible reward and the achieved reward,  $\mu_{j*} - \mu_{jk}$ . Then the cumulative regret through  $T$  periods across all  $J$  contexts is

$$\text{Regret}_T(w) = \sum_{t=1}^T \sum_{j=1}^J M_{jt} \left( \mu_{j*} - \sum_{k=1}^K w_{jkt} \mu_{jk} \right), \quad (5)$$

where  $w = \{w_{jkt} | j = 1, \dots, J; k = 1, \dots, K; t = 1, \dots, T\}$  denotes the collection of all action weights. We return to analysis of regret in terms of specific solution methods.

Regret is a frequentist criterion used in evaluating the efficiency of a method or algorithms. By quantifying regret through a finite-time period  $T$ , we can study its properties as  $T \rightarrow \infty$ . These are known as *asymptotic* properties. The particularly useful quantity of an algorithm is its asymptotic upper bound of regret (i.e., worst case scenario). This leads us to notions of an *asymptotic optimal* solution with respect to regret. Asymptotically optimal means that a policy provides strong guarantees of good performance, known as high probability of low regret (Robbins and Lai 1985; Lai 1987). Next, we address the precise formulation of the optimization problem.

### 2.1.1 Objective function

With its pieces now defined, the optimization problem can be stated as follows. The objective is to maximize the expected discounted infinite sum of rewards,

$$\begin{aligned} \max_{\{w\}} \int_{\Sigma} \int_{\bar{\beta}} \left[ \int_{\beta_1 \dots \beta_J} \mathbb{E}_f \left\{ \sum_{j=1}^J \sum_{t=1}^{\infty} \gamma^t R(w_{jt}, \beta_j) \right\} g(\beta_1 \dots \beta_J | \bar{\beta}, \Sigma) d\beta_1 \dots d\beta_J \right] p(\bar{\beta}) p(\Sigma) d\bar{\beta} d\Sigma \\ \text{s.t. } 1 = \sum_{k=1}^K w_{jkt} \forall j, t \end{aligned} \quad (6)$$

where, at each period  $t$ , we select  $\{w_t\} = \{w_{jkt} | j = 1, \dots, J; k = 1, \dots, K\}$  the collection of context-specific weight vectors such that  $1 = \sum_{k=1}^K w_{jkt}$  for each context  $j$ . Additionally,  $0 < \gamma < 1$  is a discount factor. The prior distributions characterizing parameter beliefs are  $g(\beta_j | \bar{\beta}, \Sigma)$ , with priors,  $p(\bar{\beta})$  and  $p(\Sigma)$ .

Since the optimization problem is a learning problem, it is naturally stated from a Bayesian perspective. However, at times, we will adopt a frequentist perspective for ease of exposition and because the methods can still be evaluated using frequentist criteria (e.g., regret, as discussed above).

### 2.1.2 Learning and the MAB

The essence of any MAB is the tradeoff between earning and learning. The challenge is that in order to earn (maximize the objective function) we need to learn the unknown parameters,  $\theta$ . But unlike some adaptive settings, e.g., adaptive conjoint analysis (Toubia et al. 2004; Toubia et al. 2007), parameter learning is not the ultimate goal here. Instead of adaptively maximizing precision (learning), we are interested in precision only for profit (learning while earning). In order to maximize cumulative reward (minimize cumulative regret) it is not desirable to allocate too many observations to actions that are believed to be inferior. The meanings of “too many” and “inferior” are data-driven in the MAB solutions. In fact, all MAB solutions can be framed as dynamically providing the number of observations that should be allocated to different treatments in the experiment in order to resolve the explore-exploit dilemma.

Learning means that we update our beliefs each period in light of all of the data seen up to that point, we express the data as the collection of rewards and observations for each context-action combination over time in addition to the stationary set of vectors describing action by attributes,

$$D_t = \{(y_{jk1}, m_{jk1}), \dots, (y_{jkt}, m_{jkt}) \forall j, k\} \cup \{x_1, \dots, x_K\}. \quad (7)$$

Note if  $m_{jkt} = 0$  then  $y_{jkt} = 0$  because action  $k$  was never selected in context  $j$  in time period  $t$ . We can also take subsets of this data for any context denoted by  $D_{jt}$ , which includes only data observed in context  $j$  and none of the other  $J - 1$  contexts.

Based on the data,  $D_t$ , we update our beliefs to form a joint posterior distribution,  $p(\theta|D_t)$ . For these procedures, we take as given that samples from the posterior distribution of  $\theta$  are attainable. Our focus is what to do with the posterior samples and which posterior quantities to compute.

Our learning is focused on  $\beta_j$  for each context. We leverage both the attribute and the hierarchical structures of the problem to improve learning about the parameter, i.e., reducing variance in posterior  $p(\beta_j|\bar{\beta}, \Sigma, D_t)$ .

Due to the attribute structure, within any context  $j$ , instead of learning about the ex-

pected reward,  $\mu_{jk}$ , using only observations from that action  $k$ , it is possible to leverage information from all  $K$  actions. That is, we learn the parameter  $\beta_j$ , through all actions, unlike the basic bandit problem, because of the shared attribute space.

Due to the hierarchical structure, instead of learning about each  $\beta_j$  vector by using only the data from context  $j$ , we learn by sharing information across contexts, since we can also learn about the population parameters of  $g(\bar{\beta}, \Sigma)$ . This is achieved through the partial pooling of context-specific and population-level information in a hierarchical model.

## 2.2 Framing as a Markov decision process

Conceptually, given the dynamic tradeoff between immediate reward and future reward through learning, it is natural to frame this optimization as a Markov decision process. However, we will discuss below that even typical approximate dynamic programming methods cannot overcome the curse of dimensionality here (Powell 2011). We take an action, characterized by allocation weights  $w_{jt}$  of  $d$ -dimensional vectors  $x_1, \dots, x_K$ , given a state. The state is characterized by the current beliefs about the unknown parameters. In a more Bayesian notation, this is expressed as the current posterior distribution,  $\hat{p}(\theta|D_t)$ , but we will denote this by  $\hat{\theta}_t$ , which can be interpreted as a draw from the posterior. We do this for ease of exposition although it is typically associated with a frequentist interpretation (i.e., point estimates).

The state changes as we interact with and learn more about our environment. As time progresses, we update our posterior distribution of  $\theta$  from  $\hat{\theta}_t$  to  $\hat{\theta}_{t+1}$ . These dynamics in beliefs can be characterized by the following recursive definition of value function  $V(\theta)$  satisfying the Bellman equation,

$$V(\hat{\theta}_t) = \max_{\{w_t\}} \left\{ \sum_{j=1}^J E_f R(w_{jt}, \hat{\theta}_t) + \gamma \int_{\hat{\theta}_{t+1}} V(\hat{\theta}_{t+1}) dp(\hat{\theta}_{t+1}|\hat{\theta}_t) \right\} \quad (8)$$

where  $\{w_t\}$  is the collection of all weights  $w_{jkt}$  for all  $k$  and  $j$  in period  $t$ . We also note that  $E_f R(w_{jt}, \hat{\theta}_t)$  only depends on the posterior distribution of  $\beta_j$ , but we keep  $\hat{\theta}_t$  to emphasize that it is the expected immediate reward (based on current beliefs).

While the Bellman equation looks like it could be tackled with typical methods when

written in this generality, this optimization problem is intractable due to the high dimensionality of both the state and action spaces (Powell 2011). Thus, this Bellman equation cannot be solved directly nor directly approximated as we discuss next.

### **2.2.1 Approximate dynamic programming methods cannot be used**

Approximate dynamic programming (DP) methods cannot use value iterations for the Bellman equation (8) to solve the problem of interest for various reasons. In particular, we highlight how three features, which differentiate this MAB from the basic MAB, make the approximate DP methods impractical here. Even for the basic MAB, which is associated with a Bellman equation that looks standard, the dynamic programming methods used to solve that are known to have computational complexity that is exponential in number of arms  $K$  (Brezzi and Lai 2002). And the MAB that we are aiming to solve is even more complex than the basic MAB for the following three reasons.

### **2.2.2 Impact of attribute-based structure**

First, the actions are not independent; rather they are inter-related through a set of attributes. When making an inference about action  $k$  with  $x_k$ , we are leveraging information from observations of similar actions, such as, action  $l$  as long as  $x_l$  is not orthogonal to  $x_k$ . The attribute-based nature makes this version of the bandit problem not “indexable” (Whittle 1980) since performance from playing action  $l$  can affect how we learn about action  $k$ . So the problem is not separable into  $K$  subproblems.

The attribute-based bandit problem no longer has an exact solution, but it does have an asymptotically optimal one (Dani et al. 2008). Ignoring the attribute structure has been shown to reach lower levels of reward (Rusmevichientong and Tsitsiklis 2010).

### **2.2.3 Impact of heterogeneity (hierarchical structure)**

Second, the contexts differ in unobserved ways, as the effects of the attributes in different contexts  $i$  and  $j$  can differ as well,  $\beta_i \neq \beta_j$ . We refer to this as both unobserved heterogeneity and hierarchical structure. As a result, the dimensionality of the state space (number of parameters

being learned) can be very large. Some work has shown approximations for a bandit problem with independent actions (no attributes) and discrete unobserved heterogeneity, known as the expected Gittins index (Krishnamurthy and Mickova 2010; Krishnamurthy and Wahlberg 2009), and it has been applied to marketing (Hauser et al. 2009). However, to our knowledge there is no solution that incorporates unobserved heterogeneity in the presence of an attribute-based problem. Further, the solution framework used in the (homogeneous) attribute-based bandit is not readily extendable to such a case.

#### **2.2.4 Impact of batched decisions (delayed feedback)**

Third, the firm does not select one action  $1, \dots, K$  at a time; instead, it sets the weights  $w_{jt}$  indicating what *proportion* of times each arm should be played. This is referred to as batched decision making. Therefore, the firm's decision is to select a vector from a multidimensional and continuous action space. However, such action spaces are not amenable to direct approximations.

Some work has examined the case of selecting a combination of unique actions each decision period (Sundaram 1993). However, it does not allow for selecting the same arm multiple times. Other work handles delayed feedback, which is the common interpretation of batched decision making, but it does so in an ad hoc manner (Agarwal et al. 2008; Eick 1988). For instance, Agarwal et al. (2008) use a hypothetical run technique, in which they consider what would be the fraction of times each action would have been taken (based on recalculating each action's score), if they could take actions one-by-one. Besides not being computationally efficient, this batching method depends on the form (and scale) of the scoring rule and does not explicitly handle the natural parameter uncertainty.

### **2.3 A look ahead**

Therefore, there does not exist an exact solution to this problem nor a direct approximation; instead, we must solve the problem with policies that are asymptotically optimal or based closely on policies that are asymptotically optimal.

The Gittins index is an elegant solution but its limitations have been documented, espe-

cially in the reinforcement learning literature (Auer et al. 2002; Sutton and Barto 1998). In the presence of *any* of the three key features of the managerial problem of interest (e.g., attributes, heterogeneity, batching), the Gittins index is no longer guaranteed to be optimal or even practically applicable. Given that we will be focused on problems that do relax the assumptions underlying the optimality of the Gittins index, we consider other methods that solve the basic MAB and are more amenable to extensions to the even more complex versions of the MAB.

The solution approach that we propose builds upon methods using randomized probability matching, which is grounded in well-established Bayesian methods and have been recently supported by (frequentist) asymptotic theory. While the backbone of our approach is randomized probability matching, which does not directly solve the dynamic programming problem, its theoretical performance is nearly optimal for that same optimization problem described in the dynamic programming setup above. Regret, asymptotic optimality, and the correspondence between the finite-horizon undiscounted and infinite-horizon undiscounted problems will be discussed more formally in the context of so-called upper confidence bound policies.

While the above MDP framework is useful, only one of the special cases of this optimization problem corresponds to a Bellman equation that is associated with an exactly optimal solution, and that is a very limited case, the basic MAB.

### **3 Basic bandit problem**

There is a large gap between our motivating problem and the basic multi-armed bandit problem. We continue to illustrate this gap by precisely stating the basic MAB and describing its exactly optimal solution. Recall, the basic MAB corresponds to the case when all actions are independent of one another (no shared attributes), each action's payoff distribution is identical across contexts (no unobserved heterogeneity), and the decision-maker has the hypothetical ability to take each action one-at-a-time (no batched decisions).

This is the least adorned version of the bandit problem. In the presence of an additional assumption (e.g., geometric discounting of rewards) there is an exactly optimal solution that maximizes the infinite discounted sum of rewards. This optimal solution, known as the Gittins

index policy, dynamically scores actions based on expected performance (exploitative value) and potential value from learning (exploration bonus), and selects the action with the highest index value. The method is based on Bayesian learning and an iterative solution that satisfies a Bellman equation for each action (Gittins and Jones 1974; Gittins 1979).

Consider a  $K$ -armed bandit, with independent actions (no attributes of actions, so  $x_k$  are  $K$ -dimensional indicator vectors,  $d = K$ ), one context (no unobserved heterogeneity, so  $J = 1$  and subscript  $j$  is suppressed), with sequential decision making (one-at-a-time, no batching, so  $M_t = 1$  for all  $t$ ), and geometric discounting of the stream of rewards. So we can simplify the notation introduced in Section 2. The unknown parameter vector is  $\theta = (\mu_1, \dots, \mu_K) = \mu$ , the expected rewards of each arm, where  $\mu_k = E_f(y|x_k)$  and  $Y_{kt} \sim f(y|x_k)$ . The per period reward is  $R(A_t, \mu) = Y_{kt}$ , its associated action is  $A_t \in 1, \dots, K$ . The objective is to maximize expected infinite discounted sum of rewards. Therefore,

$$\max_{A_t \in 1, \dots, K} \int_{\mu_1} \cdots \int_{\mu_K} E_f \left\{ \sum_{t=1}^{\infty} \gamma^t R(A_t, \mu) \right\} p(\mu_1) \cdots dp(\mu_K) d\mu_1 \cdots d\mu_K \quad (9)$$

because the joint prior  $p(\theta) = p(\mu_1) \cdots dp(\mu_K)$  is separable into the priors  $p(\mu_k)$  for all  $k$ , and  $0 < \gamma < 1$  is a discount factor.

### 3.1 Solving the basic bandit problem as an MDP with the Gittins index

From the objective function immediately above in the setup for the basic MAB, we explain the basic ideas behind the classic solution known as the Gittins index (Gittins 1979; Gittins and Jones 1974). When we observe action  $k$ , we do not learn about any of the other  $K - 1$  actions. This feature makes the problem separable into  $K$  separate “one-and-a-half-armed” bandit problems. This name refers to the one uncertain arm with an unknown expected reward competing against a certain arm with a known reward. This is a far simpler problem than the  $K$ -armed problem because we only need to make judgements about the expected future value of playing each uncertain arm compared to a single.

The solution comes from formulating the problem from a Bayesian perspective, as we have already done. The unknown parameters are the actions’ expected rewards,  $E_f(Y_{kt}) = \mu_k$ ,

and we quantify that uncertainty and learn by assuming a prior  $p(\mu_k)$ , for each parameter, updating it as new data arrive. The following framework can be shown in general for reward distributions that are in the one-parameter exponential family of distribution (Brezzi and Lai 2002). However, we show it only for the case of Bernoulli rewards for clarity.

For the Bernoulli case, we have reward distribution,  $f(y|\mu_k) = \text{Bernoulli}(\mu_k)$ , and prior,  $p(\mu_k) = \text{beta}(a_0, b_0)$ , for all  $k$ . Then Bayes updates are made sequentially after each Bernoulli trial. If there is a success the reward is  $y_{kt} = 1$ , otherwise  $y_{kt} = 0$ . So after period  $t$ , the beta distribution shape parameters are  $a_{kt} = a_{k0} + \sum_{\tau=1}^t y_{k\tau}$  and  $b_{kt} = b_{k0} + \sum_{\tau=1}^t (m_{k\tau} - y_{k\tau})$ , incorporating the counts the number of successes and trials for each arm  $k$ . Then the information gain (state transition) occurs with the outcome of each trial, so the state transition probabilities can be fully described by the likelihood of a successful trial based on current beliefs,

$$\Pr(Y_{kt} = 1|a_{kt}, b_{kt}) = E_{p(\mu_k)}(\mu|a_{kt}, b_{kt}) = \frac{a_{kt}}{a_{kt} + b_{kt}}. \quad (10)$$

Since the rewards are binary, the state space is described simply by pairs of non-negative integers. Each of the ‘‘one-and-a-half-armed’’ bandit problems is an optimal stopping problem (i.e., when to stop exploration and begin exploitation). This optimization problem can be described by a Bellman equation, and the value function satisfying it is

$$\begin{aligned} V(a_{kt}, b_{kt}, \gamma) &= \max \left\{ \frac{G_{kt}}{1 - \gamma}, \right. \\ &\quad \left. [1 + \gamma V(a_{kt} + 1, b_{kt}, \gamma)] \frac{a_{kt}}{a_{kt} + b_{kt}} \right. \\ &\quad \left. + [0 + \gamma V(a_{kt}, b_{kt} + 1, \gamma)] \frac{b_{kt}}{a_{kt} + b_{kt}} \right\} \\ &= \max \left\{ \frac{G_{kt}}{1 - \gamma}, \frac{a_{kt}}{a_{kt} + b_{kt}} + \gamma \left[ V(a_{kt} + 1, b_{kt}, \gamma) \frac{a_{kt}}{a_{kt} + b_{kt}} + V(a_{kt}, b_{kt} + 1, \gamma) \frac{b_{kt}}{a_{kt} + b_{kt}} \right] \right\}. \end{aligned} \quad (11)$$

For any values of  $(a_{kt}, b_{kt}, \gamma)$ , there is an exactly optimal value of  $G_{kt}$ ; for this Bernoulli  $K$ -armed bandit problem, that value of  $G_{kt}$  is the Gittins index (dynamic allocation index). This is demonstrated in the original derivation (Gittins 1979) and reviewed in various applications (e.g., Hauser et al. 2009). It can be interpreted as the present value of discounted infinite sum of

future rewards from an action while taking into account the value of reducing uncertainty. It is the exact solution to a quintessential learning problem. There are tables showing the values of the Gittins index for different values of beta distribution shape parameters and discount factors, as well as parameter values for other distributions in the exponential family (Brezzi and Lai 2002; Gittins et al. 2011).

The work of Gittins is seminal because it solved a classic sequential decision making problem previously thought to be intractable (Berry and Fristedt 1985; Gittins et al. 2011). The Gittins index itself has attracted many alternative proofs (Tsitsiklis 1986, 1994; Weber 1992). The property that it can be obtained separately for each action is the feature that makes it an *index policy* (Whittle 1980), which has been identified for a slightly broader class of problems. While Gittins index has been applied in marketing and management science (Hauser et al. 2009; Bertsimas and Mersereau 2007), these same applications note that computing the Gittins index cannot be done in closed-form, and it only solves a narrow set of problems with restrictive assumptions.

### 3.2 Approximate Gittins index

In practice, one of two methods is used: iterative algorithm to obtain the Gittins index or a well-established approximation, which we refer to as the approximate Gittins index (AGI) (Brezzi and Lai 2002, Section 2.3). Brezzi and Lai (2002) show that the AGI approximates the optimal policy (the Gittins index), maximizing the objective function for the problem. It has these properties because the AGI is an approximate solution to the underlying optimal stopping problem (i.e., when to stop experimenting with each arm).

Without describing its details, we want to draw extra attention to the structure of the approximation to the optimal solution. It can be interpreted as an option value. It is the sum of the posterior mean and an increasing function of posterior variance. That is,

$$G^{(kt)} \approx \text{AGI}^{(kt)} = E_{f_k(y|\mu)} + \sqrt{\text{Var}_{p_{kt}(\mu)}} \psi \left( \frac{\text{Var}_{p_{kt}(\mu)}}{c \text{Var}_{f_k(y|\mu)}} \right) \quad (12)$$

where  $0 < \gamma = e^{-c} < 1$  and the function  $\psi(\cdot)$  is a piecewise non-linear function discussed

and reported in Brezzi and Lai (2002, Section 3.2). We refer to that extra value above the mean as an *exploration bonus*. On the one hand, when the future is not valued at all  $\gamma \rightarrow 0$ ,  $c \rightarrow -\infty$ , and the Gittins index approaches the mean, so a purely myopic policy is used, i.e., the exploration bonus vanishes. On the other hand, when the future value of learning carries greater importance,  $\gamma \rightarrow 1$ ,  $c \rightarrow \infty$ , i.e., the exploration bonus increases.

For the beta-Bernoulli case, where  $f_k$  is Bernoulli and  $p_k$  is beta prior with  $p_{kt}$  denoting the beta posterior through  $t$  periods of data. Then the components of the AGI are as follows,

$$\begin{aligned} \mathbb{E}_{p_{kt}(\mu)} &= \frac{a_{kt}}{a_{kt} + b_{kt}} \\ \text{Var}_{p_{kt}(\mu)} &= \frac{a_{kt}b_{kt}}{(a_{kt} + b_{kt})^2(1 + a_{kt} + b_{kt})} \\ \mathbb{E}_{f_k(y|\mu)} &= \mu_k \\ \text{Var}_{f_k(y|\mu)} &= \mu_k(1 - \mu_k). \end{aligned}$$

Then the Gittins index for Bernoulli bandit,  $\text{AGI}_c(a_{kt}, b_{kt})$ , takes the form,

$$\text{AGI}_c(a, b) = \frac{a}{a + b} + \sqrt{\frac{ab}{(a + b)^2(1 + a + b)}} \psi \left( \frac{\frac{ab}{(a+b)^2(1+a+b)}}{c \left( \frac{a}{a+b} \right) \left( 1 - \frac{a}{a+b} \right)} \right) \quad (13)$$

To see the connection in the structure of the optimal solution, we note the clear connection between the Bellman equation for any arm  $k$  and the AGI for arm  $k$ . The second term is the continuation value, based on the expected future value gained from learning about the true expected reward.

However, the Gittins and AGI solve an extremely narrow set of problems. If we were to restrict ourselves to using only optimal solutions, we would restrict ourselves to a very special case of the motivating problem stated in the introduction. While one stream of literature in economics and operations research has extended the Gittins index in various direction, for instance switching costs and budget constraints (Banks and Sundaram 1994; Sundaram 1993), they remain difficult to compute and heavily reliant on the separability (statistical independence) of all actions.

The Gittins index is an elegant solution but its limitations have been documented, espe-

cially in the reinforcement learning literature (Auer et al. 2002; Sutton and Barto 1998). In the presence of *any* of the three key features of the managerial problem of interest (e.g., attributes, heterogeneity, batching), the Gittins index is no longer guaranteed to be optimal or even practically applicable. Given that we will be focused on problems that do relax the assumptions underlying the optimality of the Gittins index, we consider other methods that solve even this relatively simple bandit problem but are more amenable to extensions.

### 3.3 Beyond the Gittins index

While this basic bandit problem has an exactly optimal solution, there are two families of methods that are asymptotically optimal for the basic bandit problem: *upper confidence bound* (UCB) policies and *randomized probability matching* (RPM). These are often called heuristics or suboptimal policies, but they are important as they are easy-to-compute solutions that have been proven to perform nearly optimally in maximizing the cumulative reward for finite – yet arbitrary large – time horizons (e.g., asymptotic optimal) (Brezzi and Lai 2002; Lai 1987). In addition, the solutions can be readily extended to handle key components that comprise the class of real-world, large-scale problems we want to solve.

The intuition of this simplest UCB policy is as follows. Given the sample mean of rewards earned for each action, we are uncertain about the true mean due to sampling variation. Like the case of the Gittins index, we want to score each action. We assign a value to each action, which serves as its option value, combining the expected reward and the expected potential gain from learning about the payoff distribution for that action (e.g., if the true mean is actually higher than reflected by current beliefs). In short, the policy calls for acting optimistically in the face of uncertainty (Auer et al. 2002; Lai 1987; Lai and Robbins 1985). The values assigned to each action can be interpreted as a particular upper bound of a confidence interval for the action’s expected reward. Note that they are not exactly the frequentist confidence interval of the sample mean; instead, they are derived quantities so their size is set (and changes) in order to maximize the cumulative sum of rewards from selected actions. Nevertheless, the UCB is a frequentist idea and the many versions of UCB algorithms are proven to minimize cumulative regret with high probability asymptotically (Auer et al. 2002).

In contrast to the UCB, the intuition of the randomized probability matching principle is that the proportion of times an action is taken should be equal (i.e., matched) to the posterior probability that an action is the optimal. Naturally, RPM comes from a Bayesian perspective. In essence, the exploration/exploitation dilemma is resolved through sampling from the joint posterior distribution of expected rewards: an action is taken if it happens to have the highest expected reward among that joint sample. So the actions believed to have higher mean rewards will be likely be sampled more often than those with lower means. But an action believed to have lower mean can be taken if there is sufficient uncertainty around that value so that posterior distribution of the mean covers values high enough values, so there is some chance that in a posterior draw it seems the action in question is better than the alternatives. This idea was first proposed for basic bandit problem (Thomson 1933) and has been applied repeatedly in adaptive clinical trials (Berry 1972, 1978, 2004), with developments largely separate from the research on the Gittins index and UCB.

The intuition of RPM is simple, so it may even seem disconnected from the optimization problem at hand. That is because, unlike a Gittins index or UCB, RPM is not designed to explicitly optimize an objective function; RPM is a heuristic. Nevertheless, it is a remarkably powerful MAB solution with strong theoretical guarantees from both a Bayesian and a frequentist perspective (Agrawal and Goyal 2012; Berry and Eick 1995; Chapelle and Li 2011; Graepel et al. 2010; Scott 2010).

Both UCB and RPM can be used for the basic bandit problem. While UCB has been more popular, we will use RPM since it is more flexible for accommodating all of the features of the real-world problems that we want to solve. However, before we return to the RPM, we review the work in UCB family of algorithms, starting with the UCB policy solves the basic bandit problem.

### **3.4 Upper confidence bound (UCB) policies**

UCB policies have two major advantages relative to the Gittins index: they are easy-to-compute while still being nearly optimal policies, and they are more flexible than the Gittins index in accommodating actions described by an attribute structure. Lai and Robbins (1985) first

obtained these “nearly optimal index rules in the finite-horizon case” where the indices can be interpreted as “upper confidence bounds for the expected rewards” of the actions (Brezzi and Lai 2002, pp. 88-89). While these index rules do not provide the exactly optimal solution to the optimization problem with discounted infinite sum of expected rewards solved by the Gittins index, these rules are asymptotically optimal. This means as  $T \rightarrow \infty$ , they have nearly optimal performance with respect to maximizing the expected sum of rewards through  $T$  periods “from both the Bayesian and frequentist viewpoints for moderate and small values of  $[T]$ ” (Brezzi and Lai 2002, pp. 88-89).

Asymptotic theory links the finite-horizon undiscounted case and the infinite-horizon discounted multi-armed bandit problem (Brezzi and Lai 2002, pg. 97). As the discount factor  $\gamma \rightarrow 1$ , the UCB in Lai (1987, pg. 1113) directly approximates the Gittins index. When setting  $T = (1 - \gamma)^{-1}$ , as  $\gamma \rightarrow 1$ , then the UCB in Lai (1987) is not only asymptotically optimal for the finite-horizon undiscounted problem but also for the infinite-horizon undiscounted problem from Gittins.

These UCB policies provide the backbone of the solutions in reinforcement learning to the MAB (Agrawal 1995; Auer et al. 2002). Their structure is the sum of expected reward and an exploration bonus, just like the AGI (Brezzi and Lai 2002). For instance, in the focal algorithm in Auer et al. (2002) the UCB for arm  $k$  after  $t$  plays of all  $K$  arms is based on only the sample mean reward of the arm, where after playing the  $k$ th arm  $n_k(t)$  times, the rewards observed are  $y_{k1}, \dots, y_{kn_k(t)}$ . Given this data, the arms are scored according to

$$\text{UCB1}_k = \frac{1}{n_k(t)} \sum_{i=1}^{n_k(t)} y_{ki} + \sqrt{\frac{2 \log(t)}{n_k(t)}}, \quad (14)$$

then the arm with the highest UCB value is selected.

While the exploration bonus may seem opaque at first glance, all UCB variants are derived using regret theory. Like the regret defined earlier, in this basic MAB the finite-horizon undiscounted case regret “is the loss due to the fact that the globally optimal policy is not followed all the times” (Auer et al. 2002). Thus, when  $\mu_* = \max_k \mu_k$  and  $A_t$  is in  $\{1, \dots, K\}$ , then  $\text{Regret}_t = \sum_{\tau=1}^t (\mu_* - \mu_{A_\tau})$ .

To obtain a UCB and prove that it is asymptotically optimal, the value of information (exploration bonus) is reverse engineered to ensure that cumulative regret (the difference between the cumulative reward and the optimal cumulative reward) grows slowly at logarithmic rate in time, with arbitrarily high probability. Lai and Robbins (1985) first show that the regret for a stylized multi-armed bandit has to grow at least logarithmically in time (number of plays). The proofs are beyond the scope of this document, but we note that they rely on Chernoff's inequality to bound a sum of random variables (Auer et al. 2002).

As discussed earlier, the basic MAB is important for understanding the essential exploration/exploitation dilemma, but it is stylized, lacking of many of the key features that make the real-world MAB problems interesting yet challenging. Of the three, the feature that has attracted most attention from the MAB literature has been allowing actions to be described by attributes.

## 4 Attribute-based bandit problem

The attribute-based multi-armed bandit problem generalizes the basic MAB by allowing the actions to be interrelated instead of independent. Each action is fully described by a combination of attributes, and the set of attributes is common across actions. Unlike the basic bandit problem, when one action is selected (observed), not only do we learn about the reward distribution of the selected action but we also learn about other actions similar to the selected one. Since we assume the expected reward is a linear combination of these attributes, the problem is commonly called the “linear bandit” (Dani et al. 2008).

A closely related problem is known as the “contextual bandit” (or “bandit with side information”), in which the decision maker encounters an observed covariate and then must select an action (Langford and Zhang 2008). This “side information” case is similar to the attribute-based case that we consider since the expected reward is also assumed to be a linear combination of observed covariates and an unknown coefficient parameter vector (Langford and Zhang 2008). At their core there is a common regression structure. However, the two problems differ because, in the case of side information, the covariates are not associated with

an action, so the decision maker does not select which covariates to observe, and hence it does not resemble the adaptive sequential experiment problem we seek to solve. For this reason, we address the attribute-based MAB.

Allowing the actions to be described by a common set of attributes has been a key innovation in the UCB family of algorithms. Recent work in reinforcement learning has obtained policies, such as UCB-Lin or UCB-GLM, that are asymptotically optimal with respect to regret (Dani et al. 2008; Filippi et al. 2011, Rusmevichientong and Tsitsiklis 2010). These algorithms combine the good properties that we know from regression and classical design (generalized linear model; GLM) with the tradeoff between exploration and exploitation (UCB). Again, these UCB policies are derived from a frequentist perspective.

For any action  $k$  considered, we observe its attributes  $x_k$ , and we have a current parameter estimate of  $\beta$ , then we define the UCB-GLM. As usual, it is the sum of the predicted mean and exploration bonus, and each part incorporates the attribute-based structure,

$$\text{UCB-GLM}_{kt} = \text{link}^{-1}(x_k' \beta_t) + \|x_k\|_{Q_{t-1}^{-1}} \rho(t). \quad (15)$$

We decompose the exploration bonus in the Algorithm 1, below. As more information accumulates, the exploration bonus decreases, and the UCB-GLM converges to the predicted mean reward (similar to UCB1, AGI, and Gittins index).

The key aspects of this algorithm are its connections to linear regression. The parameter to be learned is the coefficient vector  $\beta$ , hence the algorithm relies on the frequentist properties of this regression parameter estimate (e.g., consistency).

We also note that  $\rho(t)$  is the asymptotic approximation of the upper bound of the size of the confidence area around the estimate of parameter  $\beta$  by time  $t$ , that is,  $\|\beta - \hat{\beta}_t\|_{Q_t} \leq \rho^*(t)$  (Filippi et al. 2010). The constants  $n_k(t)$  and  $n(t)$  are cumulative counts through period  $t$ .

The estimates of sampling variation around  $\beta$  enters the exploration through  $\|x_k\|_{Q_{t-1}^{-1}}$ , which is the matrix norm of any action covariate vector with respect to the inverse of  $Q_t = \sum_{\tau=1}^{t-1} x_{A_\tau} x_{A_\tau}'$ . This is analogous to the “ $(X'X)^{-1}$ ” of any linear regression.

To see how the UCB-GLM generalizes the UCB1, consider the case without attributes. If all actions were independent, then the matrix norm factor in the exploration bonus reduces to

a familiar function of sample size depending on the number of times the action in question was played. That is,  $\|x_k\|_{Q_t^{-1}} = \sqrt{x_k' Q_t^{-1} x_k} = \frac{1}{\sqrt{n_k(t)}}$ , the usual scaling by square-root of sample size.

This version of the UCB developed for the attribute-based MAB enjoys theoretical guarantees (e.g., asymptotic optimality). However, for the reasons we discuss below, we will not implement this UCB-GLM algorithm for the problem of interest.

## 5 Solving the hierarchical attribute-based batched bandit problem with randomized probability matching

### 5.1 UCB reaches its limit in a hierarchical/heterogeneous world but RPM does not

To account for unobserved heterogeneity in parameters, we need a policy more flexible than a UCB-based approach. In particular the UCB algorithms, although proven to perform well for the attribute-based MAB, rely heavily on the number of times each action is observed. The essential piece information captured in the exploration bonus is the *effective sample size* of each action, which accounts for how much has been learned directly about the action and indirectly about it through others with similar characteristics. This is achieved by using the linear regression framework. In other words, UCB methods rely on frequentist properties of the underlying model. But such calculations of effective sample size become much more complicated to measure in the hierarchical attribute-based MAB in the motivating problem.

The hierarchical attribute-based bandit generalizes both the basic MAB and attribute-based MAB by allowing for the same attribute-based structure to be replicated across many contexts, yet the solutions may be context-specific and differ in unobserved ways (heterogeneity). In particular, many real-world settings call for a model with varying parameters (intercepts and slopes) across contexts.

Finally, we add batched decisions to the hierarchical attribute-based MAB. However, batching (delayed feedback) could be added to any of the other versions of the problem, as dis-

cussed earlier. Similarly, the hierarchical structure could be present even without the attributes. That is, there could be a hierarchical bandit with independent actions within each context. Here, we focus on the MAB when all three features are present.

We accommodate the hierarchical features of the MAB by employing a multilevel/hierarchical model, which is standard throughout Bayesian statistics (Gelman and Hill 2007). The key benefit of such a model in this setting is that we can solve the attribute-based  $K$ -armed bandit problem sitting inside each of the  $J$  contexts by efficiently using all of the information available. This is done through *partial pooling* for the context-specific parameter of interest  $\beta_j$ .

A UCB-GLM could be applied by *fully pooling* all information across all contexts or in an entirely *unpooled* manner by treating each context as independent. However, neither is using the information efficiently. More precisely, if all contexts were assumed to have identical responsiveness to attributes, then for all  $j$ ,  $\beta_j = \bar{\beta}$ , a population-level parameter. This fully-pooled model is not desirable because it would ignore all possible heterogeneity across contexts.

At the other extreme, if we treated each context entirely independent of one another, we would estimate  $\beta_j$  based on the rewards from  $K$  actions observed only in context  $j$ . This is known as the *unpooled* model. This is not advantageous for various reasons. First, there may not even be enough observations to get informative posterior distributions (or reliable estimates) of  $\beta_j$  for each context. Second, we would be ignoring useful information, such as the overall (population-level average) responsiveness across all  $J$  contexts to the  $K$  actions, or even information from contexts that are estimated to be similar to one another.

Therefore, it is desirable to strike a balance between the population-level parameter  $\bar{\beta}$  and the hypothetical context-specific estimate, which would be obtained only using data from context  $j$ . This balance is achieved through *partial pooling* in the hierarchical/multilevel model with parameter heterogeneity. To our knowledge, such a hierarchical model has not been incorporated into a solution of a multi-armed bandit problem. We do this inside of a RPM algorithm because RPM accommodates a much more general set of models of reward, whereas the UCB algorithms are derived for a narrower set of models.

## 5.2 Randomized probability matching (RPM)

The principle of RPM suggests that the probability of taking an action is equal to the action's posterior probability of being the best one. This is achieved in a data-driven manner through an exact calculation of the posterior quantity, posterior sampling, or sampling from an approximation of the posterior distribution. The idea dates back to Thompson (1933), hence it is often known as Thompson Sampling. The idea has been most widely applied in adaptive clinical trials (Berry 1972, 1978, 2004), where the allocation of patients to treatment groups is a sequential decision problem that requires a tradeoff between learning treatment effects and improving patients' conditions. In fact, the basic idea behind RPM reflects the statistical (often Bayesian) perspective on the problem before the seminal work of Gittins, since the problem was focused on optimal allocation of observations in the context of sequential design of experiments and data-dependent sampling (Louis 1975; Wahrenberger et al. 1977).

The key reason why RPM works is that the exploration/exploitation dilemma is resolved through sampling from the posterior distribution of expected rewards. The policy draws a sample and selects the action with the highest expected reward in each sample. For instance, between two actions, the one with a higher posterior mean of expected reward is more likely to be sampled, but the inferior one may still be selected if there is some (non-zero) probability that its expected reward is higher than the superior action's. Much like the intuition of the UCB for the basic bandit problem, RPM is also optimistic in the face of uncertainty. By capturing all uncertainty in our current beliefs about the expected rewards, we are indeed making the exploration-exploitation tradeoff in the MAB. This is discussed further and more precisely below.

More recently, RPM has been introduced to the reinforcement learning literature, and its properties have been assessed using the (frequentist) asymptotic criterion typical in analyzing UCB-based algorithms (Agrawal and Goyal 2012; Chapelle and Li 2010; May et al. 2011). RPM achieves the goal of accumulating reward (minimizing cumulative regret), despite not being derived to optimize that (or any) objective function. In fact, Agrawal and Goyal (2012) prove that RPM policy is asymptotically optimal in minimizing regret for the basic MAB (i.e., independent actions). For the attribute-based MAB, there is not yet a proof of RPM's asymp-

otic optimality. However, RPM has been empirically shown to be superior to the UCB-GLM algorithm (Chapelle and Li 2010). Finally, there is also not yet any proof of asymptotical optimality for the hierarchical attribute-based MAB nor is there any empirical analysis. Thus, the current paper marks the first empirical analysis of the hierarchical attribute-based MAB.

While both the theoretical and empirical analyses of RPM are still limited, the benefits are clear (Graepel et al. 2010; Scott 2010). Contrasting RPM and UCB-based approaches, Scott (2010) makes the compelling argument about the relative downside of UCB-based algorithms: they are based heavily on tuning parameters, are restricted to a narrow set of models, and do not fully incorporate parameter uncertainty in the exploration bonus. UCB algorithms have tuning parameters because they are proven to be asymptotically optimal *up to a constant factor*. Those constants, however, must be selected before data are collected. In simulation studies, the algorithm is run separately for different values of the tuning parameters, and the best results are reported (Auer et al. 2010; Filippi et al. 2010). While this is standard practice for demonstrating performance via simulation, it is not suitable for real-world implementation. Hence, RPM may be the preferred approach in a variety of real-world settings.

Our contribution is that we extend RPM to accommodate a hierarchical model with unobserved heterogeneity and demonstrate its empirical performance. While previous work contrasts the performance of RPM for an attribute-based model (e.g., logistic regression) and an independent-arm model (e.g., binomial), among other comparisons (Scott 2010; Chapelle and Li 2010), that extant literature stops short of incorporating unobserved heterogeneity into RPM.

### 5.2.1 RPM with unobserved heterogeneity (hierarchical logistic regression example)

We revisit the motivating online display advertising problem framed as a MAB, and demonstrate its solution with RPM and a logistic regression with varying parameters across websites (contexts). Let the data at each time  $t$  be fully described by number of conversions per ad  $y_{jkt}$  out of impressions delivered per ad  $m_{jkt}$  for each of the  $J$  websites (contexts) and  $K$  ads (actions). The design matrix  $X = (x'_1, \dots, x'_K)$  is of size  $K \times d$ . We collectively denote all of the data we have observed through time  $t$  as,  $D_t = X \cup \{(y_{jk\tau}, m_{jk\tau}) : j = 1, \dots, J; k =$

$1, \dots, K; \tau = 1, \dots, t$ . Then we can summarize our hierarchical logistic regression with covariates and varying slopes,

$$\begin{aligned}
y_{jkt} &\sim \text{binomial}(\mu_{jk}|m_{jkt}) \\
\mu_{jk} &= \frac{\exp(x'_k \beta_j)}{\sum_{k=1}^K \exp(x'_k \beta_j)} \\
\beta_j &\sim \text{MVNormal}(\bar{\beta}, \Sigma) \forall j \\
\theta &= (\{\beta_j\}_1^J, \bar{\beta}, \Sigma) \\
x_k &= (x_{k1}, \dots, x_{kd}) \forall k
\end{aligned} \tag{16}$$

where there are also uninformative hyperpriors  $p(\beta)$  and  $p(\Sigma)$ .

Suppose that we have obtained the joint posterior distribution  $p(\beta_1, \dots, \beta_J, \bar{\beta}, \Sigma | D_t)$  via MCMC or approximate Bayesian methods. Then the posterior beliefs of  $\beta_j$  can be characterized by  $p(\beta_j | \bar{\beta}, \Sigma, D_t)$ , which in general, does not have a closed-form expression. Using the updated beliefs of the coefficient vector  $\beta_j$  and the design matrix  $X$  we easily obtain the joint predictive distribution of conversion rates (expected rewards),  $\mu_j = \mu_{j1}, \dots, \mu_{jK}$ . We then let the distribution of this  $K$ -dimensional vector be denoted by  $p(\mu_j | D_t)$ . While we will use this simplified notation, we could alternatively denote the expected reward as  $\mu_j(\theta)$  or denote the joint posterior as  $p(\mu_{j1}, \dots, \mu_{jK} | \beta_j, D_t) p(\beta_j | \bar{\beta}, \Sigma, D_t)$ . Nevertheless, it is important to note that the reward distributions of the actions are not independent; instead, even within any context,  $j$ , the reward distributions of the actions are correlated through the common set of attributes  $X$ . This is essential because the uncertainty around the expected rewards  $\mu_j$  are transformations of the uncertainty around the parameters  $\beta_j$ , the context-specific attribute coefficients.

### 5.2.2 Allocation probabilities

In order to translate the predictive distribution of  $\beta_j$  into action in each context, we apply the principle of RPM. For any context  $j$ , we let  $\mu_{j*} = \max\{\mu_{j1}, \dots, \mu_{jK}\}$ . Then we can define a

probability  $w_{jk}$  as,

$$\begin{aligned}
 w_{jk} &= \Pr(\mu_{jk} = \mu_{j*} | D_t) \\
 w_{jk} &= \int_{\mu_j} \mathbf{1}\{\mu_{jk} = \mu_{j*} | \mu_j\} p(\mu_j | D_t) d\mu_j
 \end{aligned} \tag{17}$$

where  $\mathbf{1}\{\mu_{jk} = \mu_{j*} | \mu_j\}$  is simply an indicator function of which ad has the highest conversion rate.

The key to computing this probability is conditioning on our beliefs about the vector  $\mu_j$  (or really  $\beta_j$ ) for all  $J$  contexts. Again, we note that by capturing all uncertainty in our current beliefs about the conversion rates, we are indeed explicitly balancing exploration and exploitation in the multi-armed bandit problem.

### 5.2.3 Empirically computing allocation probabilities

It is natural to compute these allocation probabilities by sampling from the predictive distribution  $p(\mu_j | D_t)$ . We can simulate  $G$  independent draws of  $\beta_j$ . Each  $\beta^{(g)}$  can be combined with the  $K \times d$  design matrix to form  $\mu_j^{(g)} = X' \beta_j^{(g)}$ . Again, conditional on the  $g$ th draw of the  $K$  predicted conversion rates, the optimal action is to select the ad with the largest predicted conversion rates,  $\mu_{j*}^{(g)} = \max\{\mu_{j1}^{(g)}, \dots, \mu_{jK}^{(g)}\}$ .

Across  $G$  draws, we approximate  $w_{jk}$  by computing the fraction of times each ad  $k$  is predicted to have the highest conversion rate.

$$w_{jk} \approx \hat{w}_{jk} = \frac{1}{G} \sum_{g=1}^G \mathbf{1}\{\mu_{jk}^{(g)} = \mu_{j*}^{(g)} | \mu_j^{(g)}\} \tag{18}$$

Since  $\hat{w}_{jk}$  is just a predictive quantity from a stationary distribution,  $\hat{w}_{jk} \rightarrow w_{jk}$  as  $G \rightarrow \infty$  by the ergodic theorem since we have a stationary Markov chain (Scott 2010).

Across all  $K$  ads, based on the weights  $\hat{w}_{jkt}$  computed from data through period  $t$  and the total number of (pre-determined) impressions across all  $K$  ads on website  $j$  in period  $t + 1$ ,

$M_{j,t+1}$ , the allocation is

$$(m_{1,j,t+1}, \dots, m_{K,j,t+1}) = (\hat{w}_{j1t}, \dots, \hat{w}_{jKt})M_{j,t+1} \quad (19)$$

As a result, we have now have the near optimal allocations of our ad impressions for each ad within each website for the next period, hence solving the problem of interest. Further details of the algorithm employed is provided in Algorithm 2.

#### 5.2.4 How does randomized probability matching balance exploration and exploitation?

To close, we review how RPM does in fact tradeoff exploration and exploitation to solve the MAB, hence linking it to the UCB methods and Gittins index discussed earlier, which were derived to optimize an objective function.

We can characterize RPM as drawing appropriate quantities from a posterior predictive distribution in the presence of remaining uncertainty. So, consider the case when there was no uncertainty remaining and we had perfect knowledge of all expected rewards  $\mu_j$  for all  $J$ . If we knew all conversion rates  $\mu_{j1}, \dots, \mu_{jK}$  in a context, then we would simply only play the winner, the one with the highest posterior mean of expected reward. Therefore, our allocation across  $K$  actions should be  $w_{jk} = 0$  for all  $k$  except that of  $\mu_{j*}$ , where  $w_{j*} = 1$ . This would be optimal no matter how close the second largest conversion rate is since we know with certainty that  $\mu_{j*}$  is maximal.

Of course, we are facing uncertainty in  $\mu_j$ . In fact, we face a great deal of uncertainty. In order to obtain each allocation probability,  $\hat{w}_{jk}$ , we need to integrate over the that uncertainty encoded by the predictive distribution  $p(\mu_{j1}, \dots, \mu_{jK} | D_t)$ .

While the Gittins index, AGI, and UCB family of policies all clearly balance exploration and exploitation by taking on an additive structure of expected reward and exploration bonus, it may be surprising that the explore/exploit tradeoff is made optimally by randomized probability matching policy. However, one can see a parallel since the UCB is achieving a frequentist analog of the stochastic (sampling) algorithm of RPM. By assigning a single number to capture to value of collecting future information about an action, the UCB quantifies the value of the uncertainty around the expected reward in terms of learning that the action of is in fact the

optimal action. That uncertainty around the expected reward can also be characterized by its full posterior distribution. Given the joint distribution of predicted means for all actions, the uncertainty about which one is in fact the optimal action can be characterized by computing the posterior probability that an each action has the highest expected reward.

The the proofs provided in Agrawal and Goyal (2012) support the preceding intuition. Although the proofs are obtained for the basic MAB, they show that the regret for RPM applied to solve the  $K$ -armed Bernoulli bandit problem using one-at-a-time actions through  $T$  periods can be bounded above by,

$$\text{Regret}_T = \left( \sum_{k=2}^K \frac{1}{\Delta_k^2} \right)^2 \log T, \quad (20)$$

up to a function of constants, where  $\Delta_k = \mu_1 - \mu_k$  and  $\mu_1 = \mu_* = \max\{\mu_1, \dots, \mu_K\}$ , without loss of generality we can assume the first arm is the unique optimal arm. These performance bounds are considered theoretical evidence that RPM is asymptotically optimal since cumulative regret grows at the slow rate of  $\log T$ .

For a detailed discussion of the details and the logic of the proofs, see Agrawal and Goyal (2012). The key calculations in the proofs are based on the calculation of random inequalities (e.g., probability that one arm’s expected reward is greater than another arm’s expected reward). The optimal expected reward would be achieved by only playing the optimal arm, receiving  $\mu_1$ , in expectation. Then the analysis focuses on the expected number of times the optimal arm is played compared to how many times other arms are played, which is common in similar proofs of finite-time regret bounds (Auer et al. 2012). According to RPM, expected number of times that the optimal arm is played is a function of the probability that current beliefs (joint posterior distribution  $p(\mu|D_t)$ ) suggest that the optimal arm’s expected reward is indeed the maximum,  $\Pr(\mu_1 > \max\{\mu_2, \dots, \mu_K\}|D_t)$ , For an empirical illustration, without the theoretical argument, see Figure 1 and Section 2.3 of Scott (2010).

## 6 Field experiment

Results are already obtained but excluded from this document.

## **7 Simulation study**

Results are already obtained but excluded from this document.

## 8 Algorithms

---

**Algorithm 1** UCB-GLM with logistic regression adapted from Filippi et al. (2010) for an attribute-based MAB

---

**Input:**  $x_k \forall k$  are  $d$ -dimensional action attribute vectors.

Play each action once from a set  $\{x_{A_1}, \dots, x_{A_d}\}$  that spans the  $d$ -dimensional attribute space.

**for**  $t = d + 1$  to  $T - 1$  **do**

Observe reward  $y_t \in \{1/0\}$  from playing action  $A_t \in \{1, \dots, K\}$

Collect all data  $D_t = D_{t-1} \cup (x_{A_t}, y_t)$

Estimate  $\beta$  with (quasi) maximum likelihood estimation.

$\mu_k = \exp(x'_k \beta) / \sum_{k=1}^K \exp(x'_k \beta) \forall k$

$\rho(t) = \sqrt{2 \log(t)}$

$Q_{t-1} = \sum_{\tau=1}^{t-1} x_{A_\tau} x'_{A_\tau}$

$\|x_k\|_{Q_{t-1}^{-1}} = \sqrt{x'_k Q_{t-1}^{-1} x_k}$

Value UCB-GLM $_k = \mu_k + \frac{1}{4} \sqrt{2 \log(t) x'_k Q_{t-1}^{-1} x_k}$

Act  $A_{t+1} = \arg \max_k \{\text{UCB-GLM}_k\}$

**end for**

---



---

**Algorithm 2** RPM with logistic regression with parameter heterogeneity and batched decisions

---

**Input:**  $x_k \forall k$  and  $M_{jt} \forall j, t$

**for**  $t = 1$  to  $t_{\text{initial}}$  **do**

$w_{jk} = 1/K$

**end for**

**for**  $t = t_{\text{initial}} + 1$  to  $T - 1$  **do**

Observe  $y_{jkt}$  successes out of  $m_{jkt}$  trials  $\forall j, k$

Collect all data  $D_t = (\sum_{\tau=1}^t y_{jk\tau}, \sum_{\tau=1}^t m_{jk\tau})$

Obtain joint posterior  $p(\beta_1, \dots, \beta_J, \bar{\beta}, \Sigma | D_t)$

**for**  $g = 1$  to  $G$  **do**

Draw  $\bar{\beta}^{(g)}, \Sigma^{(g)}$  from  $p(\bar{\beta}, \Sigma | D_t)$

Draw  $\beta_j^{(g)}$  from  $p(\beta_j | \bar{\beta}, \Sigma, D_t) \forall j$

$\mu_{jk}^{(g)} = \exp(x'_k \beta_j^{(g)}) / \sum_{k=1}^K \exp(x'_k \beta_j^{(g)}) \forall j, k$

$\mu_{j*}^{(g)} = \max\{\mu_{j1}^{(g)}, \dots, \mu_{jK}^{(g)}\}$

**end for**

$\hat{w}_{jk} = \frac{1}{G} \sum_{g=1}^G \mathbf{1}\{\mu_{jk}^{(g)} = \mu_{j*}^{(g)}\} \forall j, k$

$(m_{j,1,t+1}, \dots, m_{j,K,t+1}) \sim \text{multinomial}(\hat{w}_{j1}, \dots, \hat{w}_{jK} | M_{j,t+1})$

**end for**

---

## 9 References

Note: Extra references are included here that are referenced in the full paper but not necessarily cited in this document.

- Agarwal, Deepak. Bee-Chung Chen. Pradheep Elango. 2008. Explore/Exploit Schemes for Web Content Optimization. Yahoo Research paper series.
- Agrawal, Rajeev. 1995. Sample Mean Based Index Policies with  $O(\log n)$  Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*. 27(4), 1054-1078.
- Agrawal, Shipra. Navin Goyal. 2012. Analysis of Thompson Sampling for the multi-armed bandit problem. *Journal of Machine Learning Research Workshop and Conference Proceedings*. Conference on Learning Theory, 23, 39.1–39.26.
- Anderson, Eric. Duncan Simester. 2011. A Step-by-Step Guide to Smart Business Experiments. *Harvard Business Review*. March. 98-105.
- Audibert, Jean Yves. Remi Munos. Csaba Szepesvari. 2009. Exploration-Exploitation Trade-Off Using Variance Estimates in Multi-Armed Bandits. *Theoretical Computer Science*. 410(19). 1876-1902.
- Auer, Peter. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*. 3 397-422.
- Auer, Peter, Nicolo Cesa-Bianchi, Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*. 47. 235-256.
- Bellman, Richard. 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.
- Besanko, David, Ulrich Doraszelski, Yaroslav Kryukov, Mark Satterthwaite. 2010. Learning-by-Doing, Organizational Forgetting, and Industry Dynamics. *Econometrica*. 78(2) 453-508.
- Berry, Donald A. 1972. A Bernoulli Two-Armed Bandit. *Annals of Mathematical Statistics*. 43. 871-97.
- Berry, Donald A. 1978. Modified Two-Armed Bandit Strategies for Certain Clinical Trials. *Journal of the American Statistical Association*. 73. 339-345.
- Berry, Donald A. 2004. Bayesian Statistics and the Efficiency and Ethics of Clinical Trials. *Statistical Science*. 19(1). 175-187.
- Berry, Donald A. and Stephen G. Eick. 1995 Adaptive Assignment Versus Balanced Randomization in Clinical Trials: A Decision Analysis. *Statistics in Medicine*. 14. 231-246.
- Berry, Donald A., Bert Fristedt. 1985. *Bandit Problems*. Chapman Hall.
- Bertsimas, Dimitris. Adam J. Mersereau. 2007. Learning Approach for Interactive Marketing. *Operations Research*. 55(6) 1120-1135.
- Bitran, Gabriel R., Susana V. Mondschieen. 1996. Mailing decisions in the catalog sales industry. *Management Science*. 42(9) 1364-1381.
- Bradt, R. N., S. M. Johnson, and S. Karlin. 1956. On Sequential Designs for Maximizing the Sum of  $n$  Observations. *Annals of Mathematical Statistics*, 27(4), 1060-1074.
- Bult, Jan Roelf, Tom Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* 14(4) 378-394.
- Cassandra, A. R., M. L. Littman, N. L. Zhang. 1997. Incremental pruning: A simple fast exact method for partially observed Markov decision processes. *Proc. 13th Annual Conf. Uncertainty in Artificial Intelligence UAI-97*.
- Caro, Felipe. and Jeremie Gallien. 2007. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *Management Science*. 53(2). 276-292.
- Dani, V., T. P. Hayes, S. M. Kakade. 2008. Stochastic Linear Optimization Under Bandit

- Feedback. Conference on Learning Theory COLT.
- Daveport, Thomas H. 2009. How to Design Smart Business Experiments. *Harvard Business Review Magazine*. February. 1-9.
- Eick, Stephen G. 1988. Gittins procedures for bandits with delayed responses. *Journal of the Royal Stat. Society, Series B*.
- eMarketer. US Online Ad Spend to Close in on \$40 Billion. 12 Jan 2012. [www.emarketer.com/Article.aspx?R=](http://www.emarketer.com/Article.aspx?R=)  
Accessed 25 Apr 2012.
- Erdem, Tulin, Michael P. Keane. 1996. Decision Making under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets. *Marketing Science*, 15(1) 1-20.
- Gittins, John C., 1989. *Multi-Armed Bandit Allocation Indices*. 1st Edition. Wiley, New York.
- Gittins, John C. D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. In: Gani, J., Sarkadi, K., Vineze, I. (Eds.), *Progress in Statistics*. North-Holland, Amsterdam, pp. 241-266.
- Gittins, John C. 1979. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society Series B*. 41(2) 148-177.
- Gittins, John C. D. M. Jones. 1979. A Dynamic Allocation Index for the Discounted Multi-Armed Bandit Problem.
- Gittins, John C., Kevin Glazebrook, Richard Weber. 2011. *Multi-armed Bandit Allocation Indices*. 2nd Edition. Wiley, New York.
- Gonul, F. F., Mengze Shi. 1998. Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Science*. 44(9) 1249-1262.
- Gonul F. F., Frenkel Ter Hofstede. 2006. How to Compute Optimal Catalog Mailing Decisions. *Marketing Science*. 25(1). 65-74.
- Graepel, T., J. Q. Candela, T. Borchert, and R. Herbrich. 2010. Web-scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *ICML*, 13–20.
- Granmo, O.-C. 2010. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics (IJICC)*, 3(2), 207–234.
- Hauser, John R. Glen L. Urban. Guilherme Liberali. Michael Braun. 2009. Website Morphing. *Marketing Science*. 28(2) 202-223.
- Krishnamurthy, V., J. Mickova. 1999. Finite dimensional algorithms for the hidden Markov model multi-armed bandit problem. *IEEE International Conference of Acoustics, Speech, and Signal Processing*, Vol. 5. Institute of Electrical and Electronics Engineers, Washington, D.C., 2865-2868.
- Krishnamurthy, V. and Bo Wahlberg. 2009. Partially Observed Markov Decision Process Multiarmed. Bandits: Structural Results. *Mathematics of Operations Research*. 34(2) 287-302.
- Kumar, V., Rajkumar Venkatesan, Timothy R. Bohling, Denise Beckmann. 2008. The Power of CLV: Managing Customer Lifetime Value at IBM. *Marketing Science*. 27(4) 585-599.
- Lai, T. L. 1987. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *Annals of Statistics*. 15(3): 1091-1114.
- Lai, T. L. and Herbert Robbins. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6(1) 4-22.
- Langford, John and Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems*, 20,

817–824.

- Lewis, Michael. 2005. "A Dynamic Programming Approach to Customer Relationship Pricing," *Management Science*. 51(6) 981-994.
- May, Benedict C., Nathan Korda, Anthony Lee, David S. Leslie. 2011. *Optimistic Bayesian Sampling in Contextual Bandit Problems*. Technical Report 11:01, Department of Mathematics, University of Bristol.
- Mersereau, Adam J. Paat Rusmevichientong. John N. Tsitsiklis. 2009. A Structured Multi-armed Bandit Problem and the Greedy Policy. *IEEE Transactions on Automatic Control*. 54(12).
- Montoya, Ricardo, Oded Netzer, Kamel Jedidi. 2010. Dynamic Allocation of Pharmaceutical Detailing and Sampling for Long-Term Profitability. *Marketing Science* . 29( 5) 909-924.
- Powell, Warren B. 2011. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley.
- Rubin, Donald. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, Vol. 66, No.5, (1974), pp. 689
- Rusmevichientong, Paat and John N. Tsitsiklis. 2010. Linearly Parameterized Bandits. *Mathematics of Operations Research*. 35 (2). 395-411.
- Scott, Steven L. 2010. A Modern Bayesian Look at the Multi-Armed Bandit. *Applied Stochastic Models Business and Industry*. 26. 639-658.
- Sun, Baohong. Shibo Li. Catherine Zhou. 2006. "Adaptive" Learning and "Proactive" Customer Relationship Management. *Journal of Interactive Marketing*. 20(3-4) 82-96.
- Thompson, Walter R. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*. 25 285-294.
- Toubia, Olivier, John R Hauser, Duncan Simester (2004). Polyhedral methods for adaptive choice based conjoint analysis. *Journal of Marketing Research*, 41(1), 116-131.
- Tsitsiklis, John N., 1986. A Lemma on the Multi-Armed Bandit Problem. *IEEE Trans. Automat. Control*. 31, 576-577.
- Tsitsiklis, John N., 1994. A Short Proof of the Gittins Index Theorem. *The Annals of Applied Probability*. 4(1), 194-199.
- Urban, Glen L. Guilherme Liberali. Erin MacDonald. Robert Bordley. John R. Hauser. 2012. *Morphing Banner Advertisements*. Working Paper.
- Wahrenberger, David L., Charles E. Antle. Lawrence A. Klimko. 1977. Bayesian Rules for the Two-armed Bandit Problem. *Biometrika*. 64 (1). 1724.
- Whittle, P. 1980. Multi-armed Bandits and the Gittins Index. *Journal of Royal Statistical Society B*. 42(2), 143-149.
- Whittle, P. 1979. Discussion on: "Bandit processes and dynamic allocation indices" *Journal of Royal Statistical Society B*. 41, 165.
- Wind, Jerry (Yoram). 2007. *Marketing by Experiment*. *Marketing Research*. Spring. 10-16.