Bandits for Taxonomies: A Model-based Approach

Sandeep Pandey Deepak Agarwal Deepayan Chakrabarti Vanja Josifovski
Yahoo! Research, Sunnyvale, CA
{spandey, dagarwal, deepay, vanjaj}@yahoo-inc.com

January 18, 2007

Abstract

We consider a novel problem of learning an optimal matching, in an online fashion, between two feature spaces that are organized as taxonomies. We formulate this as a multi-armed bandit problem where the arms of the bandit are dependent due to the structure induced by the taxonomies. We then propose a multi-stage hierarchical allocation scheme that improves the explore/exploit properties of the classical multi-armed bandit policies in this scenario. In particular, our scheme uses the taxonomy structure and performs shrinkage estimation in a Bayesian framework to exploit dependencies among the arms, thereby enhancing exploration without losing efficiency on short term exploitation. We prove that our scheme asymptotically converges to the optimal matching. We conduct extensive experiments on real data to illustrate the efficacy of our scheme in practice.

1 Introduction

Consider an experiment involving two sets $\mathcal{S}$ and $\mathcal{A}$. Random elements of $\mathcal{S}$ that arrive sequentially are to be matched to elements in $\mathcal{A}$. Every match receives a stochastic reward with an unknown probability, and the goal is to maximize expected reward accumulated through time. In particular, we focus on maximizing the expected reward when cardinalities of both $\mathcal{S}$ and $\mathcal{A}$ are large.

Such problems are commonplace and arise in several applications (see [10] for an overview). Examples include a) product recommendation for users visiting an e-commerce website like *amazon.com* based on their demographics, previous purchase history, etc. Here, set $\mathcal{S}$ may consist of unique visitors who are matched to a set $\mathcal{A}$ of products with an objective of maximizing total sales revenue; b) a search engine placing appropriate ads on web-pages to maximize total revenue from user clicks. This problem, referred to as *content match* hereafter, motivated the current research. A similar problem but under different constraints has been studied in the context of banner advertising in [14].

When placing ads on pages in the context of content match, information that may be useful includes page attributes (e.g., page topic, content, etc.), ad attributes (e.g., theme of the ad, anchor text, landing page, etc.), and other contextual information (user demographics, their recent behavior, etc.). Assuming both pages and ads have been mapped to appropriate feature spaces (that are high dimensional) and each click on an ad earns some revenue, the general problem is to map points in a feature space of page attributes to another feature space of ad attributes after accounting for context attributes (we ignore context in subsequent discussions but discuss potential solutions in Section 6) to maximize total expected revenue. This involves exploration and exploitation at some level. We explore different ads to find good ones more effectively and we exploit the ones that are currently known to have good click rates. However, designing effective explore/exploit policies in this context is a daunting task for several reasons:

- *Data sparsity:* The feature spaces are extremely large (billions of pages, millions of ads with a lot of diversity and heterogeneity in both pages and ads) and the data extremely *sparse* since we observe only a few interactions for a majority of page-ad feature pairs.
- *Rarity of clicks:* The click-through rate (CTR hereafter) defined as the number of clicks per impression (number of showings) for a majority of page-ad feature pairs are small, leading to increased learning time.
- *Good short term performance:* Business considerations constrain us to learn CTR values in a reasonable time horizon and without incurring large drops in revenue (even in the short run). A policy that does excessive exploration provides gradual but slow revenue growth before it converges to the optimal matching. On the other hand, a policy that merely tries to achieve optimality quickly may incur an unnecessarily large revenue loss during the learning period. An ideal policy would converge rapidly to the optimal matching while having a smooth revenue profile.
- *Finite inventory:* Our policy should learn CTR values in an online setting for a large majority of page-ad feature pairs. This is important since the available inventory

is finite. For instance, we may run out of best ads for certain pages and may want to consider other lucrative opportunities. Hence, given two policies that have similar revenue profiles, we prefer the one whose CTR estimates have lower mean squared error.

To deal with the difficulties mentioned above, reducing dimensionality is of paramount importance. One approach is to assume that CTRs are simple functions of both page and ad features [15]. Such an approach is efficient if the assumed relationship is approximately true. In content match, the assumption of linearity and additivity of page and ad features is violated and leads to CTR estimates that are biased. Interactions among features are present and are extremely important for learning CTRs. Another approach is to reduce the dimension of both page and ad feature spaces and conduct learning at a coarser resolution. For example, we could work with page and ad clusters found by unsupervised learning. Such data driven clusters are useful in several contexts but may become hard to interpret and difficult to explain.

For content match, there exist *taxonomies* for both pages and ads that have been created, refined, and are well understood and routinely used by domain specialists. The existence of taxonomies simplifies the learning task by reducing the dimension of feature spaces. The reduction is accomplished by classifying ads and pages to appropriate nodes of the respective taxonomies using supervised learning. In fact, the taxonomies provide information on broad contextual themes that may lead to useful insights about the process at a macro level. This may be useful for planning and decision making. For instance, learning the presence of high CTR when electronic ads are placed on sports pages can potentially discover a lucrative market that was unknown to the business. In this paper, we discuss learning strategies using such taxonomies under the following assumptions:

- There exist classifiers that accurately classify pages and ads into their respective taxonomies. Specifically, we assume that our classifier returns a best and unique path in the taxonomy to the leaf node to which a page (or ad) belongs. Building classifiers for hierarchical content classification is an active research area (see [8] and references therein).

- We restrict our learning strategies to the last two levels of the taxonomies; potential strategies to generalize our method to entire taxonomies will be discussed in Section 6 with a detailed investigation deferred to a subsequent paper. The *key* assumption is that *dependencies are induced by lineages*, i.e., CTRs of children sharing the same parent are assumed to be related.

- In the learning process, every new arrival in $\mathcal{S}$ is matched to a single best element in $\mathcal{A}$. Precise definitions of $\mathcal{S}$ and $\mathcal{A}$ for content match is provided in Section 2.2.

1.1 Our contributions. We introduce a novel learning problem of matching feature spaces that are organized hierarchically. We formulate this as a bandit problem and propose a policy that performs better than existing bandit policies desgined for flat feature spaces. The taxonomies induce dependencies among arms of the bandit which our policy exploits in two ways: (a) it enhances exploration with a multi-stage allocation scheme that matches parents followed by a match among their children, (b) it improves estimation of rewards through *shrinkage estimation* in a Bayesian framework. We conduct extensive experiments on real data to illustrate the efficacy of our policy. In addition, we prove that our policy asymptotically converges to the optimal matching.

The roadmap is as follows. Section 2 provides an overview followed by detailed description of our policy in Section 3. In Section 4, we analyze data and derive parameter estimates that are used for experiments in Section 4.3. A survey of related work is provided in Section 5, we end with a discussion in Section 6.

2 Overview

We propose a solution to the online matching problem by combining a bandit formulation with Bayesian shrinkage estimation, to combat data sparsity and also to model dependencies induced by the reward structure. Before describing our proposed approach, we provide some background on these topics. We first describe classical multi-armed bandits. Then, we describe the bandit formulation of our problem, and associated terminology. We then discuss Bayesian shrinkage estimation in these settings. Finally, we give a high-level overview of our entire method.

2.1 The classical multi-armed bandit problem We begin by providing a high level overview of the multi-armed bandit problem and establish connection to the learning task considered in this paper for content match. The *multi-armed bandit problem* derives its name from an imagined slot machine with $k(\geq 2)$ arms. The i^{th} arm has a payoff probability p_i which is unknown. When arm i is pulled, the player wins a unit reward with payoff probability p_i. The objective is to construct N successive pulls of the slot machines to maximize the total expected reward. This gives rise to the familiar explore/exploit dilemma where on one hand one would like to gather information on the unknown payoff probabilities, while on the other hand one would like to sample arms with the best payoff probabilities, empirically estimated so far. A bandit policy or allocation rule is an adaptive sampling process that provides a mechanism to select an arm at any given time instant based on all previous pulls and their outcomes.

A popular metric to measure performance of a policy is

called *regret*, which is the difference between the expected reward obtained by playing the best arm and the expected reward given by the policy under consideration. A large body of bandit literature has considered the problem of constructing policies that achieve tight upper bounds on regret as a function of the time horizon N (total number of pulls) for all possible values of the payoff probabilities. The seminal work of [20] showed how to construct policies for which the regret is of $O(\log N)$ asymptotically for all values of payoff probabilities. They further proved that asymptotic lower bounds for the regret is also $\Omega(\log N)$ and constructed policies that actually attain them. Subsequent work has constructed policies that are simpler and achieve the logarithmic bound uniformly rather than asymptotically (see [16] and references therein). The main idea in all these policies is to associate with each arm a priority function which is a sum of the current empirical payoff probability estimate plus a factor that depends on the estimated variability. Sampling the arm with the highest priority at any point in time, one explores arms with little information and exploits arms which are known to be good based on accumulated empirical evidence. With increasing N, the sampling variability reduces and one ends up converging to the optimal arm. This clearly shows the importance of the result proved by [20] which proves that one cannot construct the variance adjustment factor to make the regret better than $\Omega(\log N)$, thereby providing a benchmark for evaluating policies.

Two policies that both have $O(\log N)$ regret might involve different constants in the bounds and may behave differently in real applications, especially when considering short term behavior. One way of comparing short term behavior of policies that are otherwise optimal in the asymptotic sense is by taking recourse to simulation experiments. This is the approach we take in this paper for evaluating our policy. One can also evaluate short term behavior by proving finite sample properties of policies but this may become extremely hard to derive except in simple situations. The main difficulty is caused by the presence of dependencies in the sampling paths.

Another approach pursued in the multi-armed bandit literature is that of devising policies that maximize the expected discounted reward, obtained by *geometrically* discounting all future rewards by a constant factor $\beta \in (0, 1)$. The optimal policy in this scenario was shown to be the "index rule" that chooses at each stage the arm with the largest "dynamic allocation index" (DAI) [4]. While this formulation may be useful in certain situations, we focus on undiscounted finite-time rewards in this paper because we are primarily interested in short term behavior.

2.2 A bandit formulation of content match. As mentioned in Section 1, we assume pages and ads have been classified into page and ad taxonomies respectively. Recall that we restrict ourselves to two successive levels of the taxonomies. We refer to the lower level nodes of the page and ad taxonomies as *page-classes* and *ad-classes* respectively; denote these by $\mathcal{S} = \{s_1, \cdots, s_u\}$ for page-classes and $\mathcal{A} = \{a_1, \cdots, a_v\}$ for ad-classes. All page-classes (ad-classes) that are children of the same parent node from the upper level are said to constitute a *page-class group* (*ad-class group*). Define the page-ad *connection* matrix $\mathcal{C} = \mathcal{S} \times \mathcal{A}$, each of whose cells has a CTR value for the corresponding (page-class, ad-class) pair (see Figure 1). Our goal is to learn matrix $\mathcal{C}$ so as to maximize the expected total number of clicks [1].

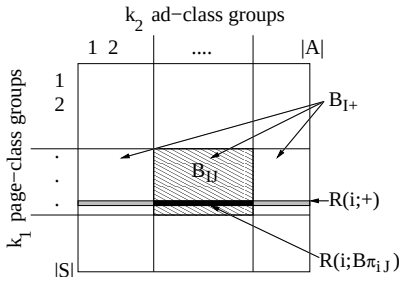


Figure 1: *Notations*

This can be formulated as a multi-armed bandit problem as follows. For each page-class we create a $v(= |\mathcal{A}|)$-armed bandit, where the arms are given by the ad-classes and the payoff probabilities by the CTR values. Thus, we have $u(= |\mathcal{S}|)$ such bandit problems that arise simultaneously.

However, the arms of each bandit and the bandits themselves are not independent of each other since $\mathcal{S}$ and $\mathcal{A}$ are partitioned into page-class and ad-class groups. In particular, the arms in the same group are likely to have similar payoff probabilities. In other words, each *block* of Figure 1 is expected to be relatively homogeneous in terms of CTR values. We exploit this structure to construct bandit policies that are optimal asymptotically and yet achieve better performance in the short run.

Next we introduce our terminology and notation. Throughout the paper, we shall use terminology closely tied to content match (page-class, ad-class, CTR matrix) to describe our problem but note that our methods are general and apply to all learning problems with a hierarchical reward structure. We also note that different terminologies are used to describe these components in the literature. For instance, in the bandit literature, ad-classes are referred to as arms and the CTRs are referred to as payoff probabilities. In the reinforcement learning literature, page-classes are called states, ad-classes are called actions and CTRs are referred to as the

[1] The general case of incorporating the auction and pricing mechanism in the learning process is beyond the scope of this paper

reward distribution. We deliberately choose this presentation style to emphasize and closely connect the mathematical problem to the motivating application of content match.

Let suffix ij denote the *cell* in $\mathcal{C}$ corresponding to page-class s_i and ad-class a_j. Let π_i and κ_j denote group ids of page-class group s_i and ad-class group a_j respectively while $B_{\pi_i\kappa_j}$ denotes the block that contains the ij^{th} cell. In particular, B_{IJ} denotes the block containing cells obtained by taking cross-product of page-classes in page-class group I and ad-class group J. Let k_1 and k_2 denote the number of page-class groups and ad-class groups respectively. Define $B_{I+} = \cup_{J=1}^{k_2} B_{IJ}$; $B_{+J} = \cup_{I=1}^{k_1} B_{IJ}$. The row for page-class s_i in matrix $\mathcal{C}$ intersecting the block $B_{\pi_i J}$ is denoted by $R(i; B_{\pi_i J})$, and $R(i; +) = \cup_{J=1}^{k_2} R(i; B_{\pi_i J})$ (see Figure 1).

For any set U of cells, let p_U, S_U and N_U denote the true CTR, number of clicks and sample size (number of impressions or pulls) after the n^{th} allocation has been made. Also, let $\hat{p}_U = S_U/N_U$ denote the maximum likelihood estimate of p_U and $CV_U = \sqrt{\frac{1-\hat{p}_U}{N_U \hat{p}_U}}$ denote the estimated coefficient of variation for U (assuming a binomial distribution with uniform CTR for cells in U). Also define $CV_{\pi_i(r)}$ be the estimated coefficient of variation with rank r (smaller CVs have higher ranks) among all blocks $B_{\pi_i J}; J = 1, \cdots, k_2$.

2.3 Dealing with Sparsity through Shrinkage Estimation. A key problem we face is the sparsity of data for a large fraction of cells. However, a small fraction of cells have relatively higher CTRs. This provides an ideal situation for improving overall estimation accuracy by using Bayesian smoothing or shrinkage estimation. The method assumes that the CTR values p_{ij}s are drawn from a prior distribution $F(\{p_{ij}\}; \theta)$ that depends on the parameter vector θ (to be estimated from data). The posterior distribution of p_{ij}s provide "smooth" estimates with better mean squared error compared to a simple scheme like maximum likelihood estimation under the independence assumption. However, the degree of smoothing depends on the choice of F. In our case, the presence of blocks B_{IJ} derived from the taxonomy motivates a separate prior for each block.[2]

2.4 High-level description of the proposed method. Since better estimation depends critically on being able to estimate the block priors, we propose a multi-stage allocation strategy that runs a bandit at the block-level (i.e., on the k_2 distinct sets $B_{\pi_i J}$ for a given page-class s_i), followed by a cell level bandit on $R(i; B_{\pi_i J^*})$ (J^* corresponds to the block selected at the first stage). The block-level bandit ensures that we explore each block often enough to estimate its prior quickly. However, since it aggregates clicks over all rows of a block, it has the potential problem of missing out on good cells in certain rows in the long run. To circumvent this, we provide a mechanism whereby our strategy switches from a block level bandit to a row-level bandit on $\{R(i, B_{\pi_i J})\}$s at some point. The switch occurs through a statistical criterion which ensures that our policy asymptotically converges to the optimal.

We describe our multi-level policy in detail next.

3 Proposed Multi-level Policy

When the n^{th} page-class arrives, we proceed as follows.

- *Estimation step:* CTR values are estimated after taking into account the outcomes of all $n-1$ previous allocations[3].
- *Allocation step:* The n^{th} page-class is matched to an appropriate ad-class based on the current estimates of the CTR values.

A detailed description follows.

3.1 Estimation Step. We assume that the number of clicks S_{ij} are binomially distributed, i.e. $S_{ij}|p_{ij} \sim Bin(N_{ij}, p_{ij})$($X|Y$ denotes the conditional distribution of X given Y throughout the paper), where N_{ij} is the total number of observations (henceforth, sample size) in cell ij, and p_{ij} is its true CTR. We assume that all S_{ij}s are conditionally independent given p_{ij}s. If the N_{ij}s are large, one can estimate the true CTRs for cells using maximum likelihood estimators (MLE) $\hat{p}_{ij} = S_{ij}/N_{ij}$. Although we know some (page-class, ad-class) pairs go well together (e.g., ski ads go well with pages about winter sports), a majority of cells would have low CTRs and hence would receive relatively fewer pulls by the bandit policy, leading to small sample sizes in the corresponding cells. Clearly, large sample size implies better information about a cell's CTR. This leads to the consideration of a shrinkage estimator, in which the estimate of a particular cell is a convex combination of a global estimator and an estimator (usually the MLE) exclusively derived from the cell information. If the MLE is based on small sample size, more weight is given to the global estimator and vice versa.

An empirical Bayes approach based on a betabinomial model provides an attractive way to accomplish shrinkage estimation in our problem setting. In particular, we assume $\{p_{ij} : ij \in B_{IJ}\}$ are drawn from a beta distribution with parameters $\alpha_{B_{IJ}}$ (mean) and $\gamma_{B_{IJ}}$ (effective sample size) which in turn induce independent betabinomial models for each block. In Section 4, we justify this choice by a retrospective analysis of existing data.

[2] smoothing across blocks can be introduced through hyperpriors on block priors but this is not pursued here

[3] For content match, clicks on ads are observed without latency.

The betabinomial model. We summarize the main properties of a betabinomial distribution here and refer the reader to [11] for complete details. This distribution arises naturally in a hierarchical Bayesian context as follows. For a single data point $\{S, N\}$, if $S|p \sim Bin(N, p)$ and $p \sim Beta(\gamma\alpha, \gamma(1-\alpha))$, the marginal distribution of S has a closed form expression and is said to be a betabinomial distribution. By Bayes theorem, $p|S \sim Beta(\gamma\alpha + S, \gamma(1-\alpha) + N - S)$ and hence the posterior mean is given by

$$E(p|S, \gamma, \alpha) = w\alpha + (1-w)(S/N) \tag{3.1}$$

where $w = \gamma/(\gamma + N)$. Note that $w \to 0$ if and only if $\gamma/N \to 0$ and corresponds to the case of "no shrinkage". For small N, w is close to 1, shrinking the posterior mean towards the global mean α. Thus, γ determines the weight attached to the prior mean α and hence the amount of shrinkage. One can also interpret γ as the *effective sample size* available a-priori. This becomes evident from the density of the beta distribution which is proportional to a binomial density with $\gamma\alpha - 1$ successes and $\gamma(1-\alpha) - 1$ failures. In practice, the parameters of the beta prior will not be known and have to be estimated from data. However, this is not possible unless we have a set of data points $\{S_k, N_k\}_k$ such that $S_k|p_k \sim Bin(N_k, p_k)$, and $p_k \sim Beta(\gamma\alpha, \gamma(1-\alpha))$. One can then estimate α and γ based on a betabinomial likelihood using maximum likelihood and hence provide estimates of the posterior distribution of p_ks. In fact, maximum likelihood estimation of α and γ have been studied in the literature and it is well known that the estimation of α is more stable compared to that of γ. In particular, estimation of γ becomes unstable if $\gamma > 3000$[5].

It is instructive to look at the mean and variance of S_k after marginalizing over p_k. In fact,

$$\begin{aligned} E(S_k) &= N_k\alpha \\ Var(S_k) &= N_k\alpha(1-\alpha)[1 + (N_k - 1)/(\gamma + 1)] \end{aligned} \tag{3.2}$$

When compared to the variance of a binomial model with parameters N_k and α, the variance term in Equation 3.2 involves an additional factor which is a function of γ. This accounts for the extra-binomial variation or overdispersion which is present in our data (see Section 4 for a detailed analysis).

Let s_i be the n^{th} arriving page-class, and suppose we allocate it to an ad-class a_j (based on a chosen policy) resulting in a click or no-click. We update the CTR values in all cells of block $B_{\pi_i\kappa_j}$ using the following scheme: we fit a betabinomial model to the block. If the fit is satisfactory, we use the betabinomial estimates for the CTRs of all cells in the block, as explained in Equation 3.1. However, if the betabinomial does not provide a good fit, we use the maximum likelihood estimates.

3.2 Allocation Step. Given an arriving page-class s_i, our multi-level policy runs bandits at two levels: first, it runs a bandit over blocks $B_{\pi_i J}$ ($J = 1 \ldots k_2$) to select a good ad-class group J^*, and then it runs a bandit over all cells in row $R(i; B_{\pi_i J^*})$ to select a good ad-class in J^*. Intuitively, the first stage can quickly identify blocks with good CTR values, since there are only k_2 of these for each s_i. This helps in focusing the search for good cells early on towards the good blocks. Also, it ensures that no block is neglected and that all block priors, critical for the estimation step, can be computed quickly.

However, if a good cell lies in a row s_i which arrives infrequently, the block estimates will be overwhelmed by other rows in the same block. If these rows have poor CTRs, this may lead the first-stage bandit to falsely believe that the entire block is poor in terms of CTR. To circumvent this problem, the first stage of our *Multi-level* policy switches from a block-based bandit to a row-based bandit. A statistical criterion based on $CV_{\pi_i(r)}$ replaces the block-level bandit in the first stage to a row-level bandit. Figure 2 provide the details.

While the *Multi-level* policy can use any multi-armed bandit as a subroutine, we use the *UCB1* scheme of [16] (see Figure 3). In it, the priorities of the arms are obtained by superimposing estimated CTRs with a component that denotes the size of an upper one-sided confidence interval containing the true CTR with overwhelming probability. The first component helps in exploiting good ad-classes while the second component supports exploration. This policy has a logarithmic regret uniformly in the number of pulls.

The CTR estimates used at the second stage of our *Multi-level* policy are derived from the beta-binomial model (if the model fits for the block). In particular, the CTR estimates are taken to be the posterior mean, and sample sizes are adjusted by adding the effective sample size parameter from the beta prior. As discussed before, the key idea is to estimate the priors quickly using the first stage, especially in the beginning when we have small samples. This provides better estimates of the individual cell CTRs by incorporating the taxonomy in the estimation through a hierarchical Bayesian model. If the betabinomial model does not fit, maximum likelihood estimates are used.

We prove asymptotic consistency of policy described in Figure 2. To be precise, we prove that asymptotically our policy converges to the optimal solution, i.e, for any given page-class, it will pull the ad-class with maximum CTR.

THEOREM 3.1. *Policy* Multi-level *is asymptotically optimal for each* $\tau(> 0)$ *and* $r(\geq 1)$.

Proof. Proved in the appendix.

4 Data Analysis and Experiments

Our current content match system uses policies that are different from the ones proposed in this paper. To perform

POLICY 1. (*Multi-level*(τ, r))

Parameters	*Select parameters* $\tau(>0)$ *and* $r(\geq 1)$. *We experiment with* $(\tau, r) = (.1, 1)$ *(aggressive switching),* $(.1, \lfloor 0.5k_2 \rfloor + 1)$ *(average switching) and* $(.1, \lfloor 0.9k_2 \rfloor + 1)$ *(conservative switching).*
Initialization	*Take one observation from each block.*
First Stage	*With the arrival of page-class* s_i*, select the appropriate block or row* J^* *by running a block or row level bandit policy. In particular,* $$J^* = \begin{cases} argmax_{J\in\{1,\cdots,k_2\}} \left(\hat{p}_{R(i;B_{\pi_i J})} + \sqrt{\frac{2lnN_{R(i;+)}}{N_{R(i;B_{\pi_i J})}}}\right) & \text{if } CV_{\pi_i(r)} \leq \tau \quad \text{[Row-level]} \\ argmax_{J\in\{1,\cdots,k_2\}} \left(\hat{p}_{B_{\pi_i J}} + \sqrt{\frac{2lnN_{B(\pi_i,+)}}{N_{B_{\pi_i J}}}}\right) & \text{otherwise} \quad \text{[Block-level]} \end{cases}$$
Second Stage	*Select the best ad-class* k^* *by running a cell level bandit on the cells of* $R(i; B_{\pi_i J^*})$. *In fact,* $k^* = argmax_{k\in R(i;B_{\pi_i J^*})} \left(\tilde{p}_{ik} + \sqrt{\frac{2ln(N_{R(i;B_{iJ^*})} + \gamma_{R(i;B_{iJ^*})})}{(N_{ik}+\gamma_{ik})}}\right)$. *Here,* $\tilde{p}_{ik}$ *is the estimated cell CTR estimated based on the model in* $B_{\pi_i J^*}$. γ_U *is the estimated effective sample size for set* U. *Hence,* $\gamma_U = \hat{\gamma}|U|$ *when the* betabinomial *model fits the block* $B_{\pi_i J^*}$*, and* $\gamma_U = 0$ *otherwise.*

Figure 2: The *Multi-level* policy

POLICY 2. (*UCB1)*

Select the best ad-class k^* *corresponding to a page-class* s_i *as follows:*

$$k^* = argmax_{k\in R(i;+)} \left(\hat{p}_{ik} + \sqrt{\frac{2ln(N_{R(i;+)})}{N_{ik}}}\right).$$

Figure 3: *UCB1* policy

a complete evaluation of alternative policies, one needs to perform tests on a live system. Due to the proprietary nature of our data, the costs involved in conducting experiments with several methods, and other additional complexities (e.g, constraints imposed by business rules), we are unable to report such results. Instead, we perform simulation studies whose designs are based on the statistical properties of our data. In other words, we simulate synthetic data using the actual structure of our taxonomies and derive all parameters needed in simulations from real data obtained from the current content match system. The data analysis presented subsequently serves two purposes:

- It proves that the assumptions underlying the betabinomial generative model are reasonable; and
- It is used to derive parameter estimates that are used in our simulations (after certain transformations on the inferred parameter values, for reasons of data confidentiality).

Using these parameter estimates, we perform an extensive simulation study to compare the policies described previously in Section 3. In particular, we show that:

- Our *Multi-level* policy has significantly better short-term performance than *UCB1*.
- Shrinkage estimation using the betabinomial model leads to better estimates of CTR values over the entire (page-class, ad-class) matrix, as measured by total mean squared error. Recall that this is important in finite inventory settings, where we may run out of the best ads for certain pages. Moreover, this increased accuracy in CTR estimation comes without any concomitant decrease in revenue.

Next, we describe the structures of the page and ad taxonomies. Then, we find the betabinomial parameters by a retrospective analysis of available data. Finally, we describe our experiments under these parameter settings.

4.1 Taxonomy structure. Both page and ad taxonomies are identical, and this common taxonomy is a tree consisting of seven levels with approximately $7K$ leaf nodes. Labeling the root as depth 0, there are 20 nodes at depth 1 and 221 at depth 2. We run our experiments on these two levels. The distribution of the number of children for nodes at depth 1 has high variance (Figure 4). This is expected, as the

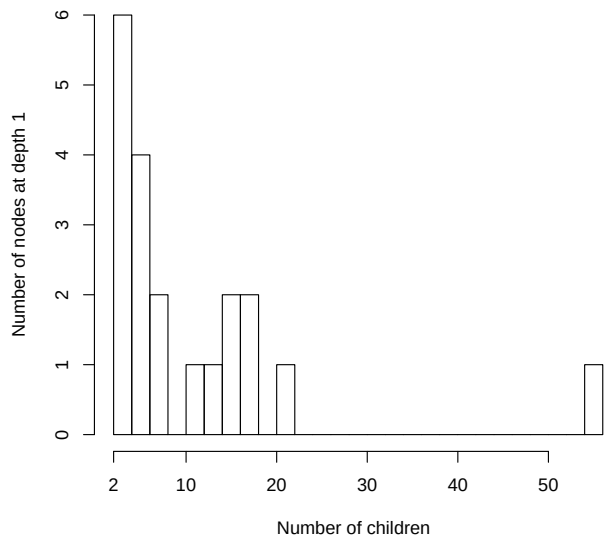


Figure 4: *Distribution of number of children for depth one nodes*

taxonomies were created manually and hence some broad themes are split up early into several sub-categories while the splitting occurs more gradually for others. Considering nodes at depth 2, our connection matrix $\mathcal{C}$ is 221×221. There are 20 groupings for both page and ad classes formed by nodes at depth 1 (ie $k_1 = k_2 = 20$) giving $20 \times 20 = 400$ blocks, i.e., $400 B_{IJ}$s.

4.2 Fitting the betabinomial. Next, we provide evidence of the appropriateness of the generative model assumptions made in this paper. To test for extra-binomial variation that motivates the betabinomial model, we analyzed click data [4] aggregated to depth 2 in our taxonomies. In particular, for any given block $B_{IJ}(I = 1 \cdots 20; J = 1 \cdots 20)$, we compute the MLE $\hat{p_{ij}} = S_{ij}/N_{ij}$ for the CTR p_{ij} of the ij^{th} cell, and get rough estimates of the block mean $\alpha_{B_{IJ}}$ and effective sample size $\gamma_{B_{IJ}}$ by fitting a beta distribution to the $\hat{p_{ij}}$s $\in B_{IJ}$. Figure 5 shows the histogram of estimated effective sample sizes $\gamma_{B_{IJ}}$s. There is wide variation in the distribution. There are several blocks with significant overdispersion, and a certain fraction of blocks having large effective sample sizes (ie, large γ) and hence little overdispersion. However, we note that the distribution results are for data obtained from the current allocation scheme used in content match for which there exist only a small fraction of cells with moderate number of pulls. Our *Multi-level* policy is meant to provide more flexibility and explore larger percentage of cells more often and hence expected to yield data that has higher overdispersion.

To provide further insights into overdispersion present in our data, we study the relationship between number of pulls in cells and Pearson residuals obtained relative to a binomial model which assumes that all CTRs in a block are identical: $p_{ij} = p_{B_{\pi_i \kappa_j}}$, with $p_{B_{\pi_i \kappa_j}}$ being estimated as $\hat{p_{B_{\pi_i \kappa_j}}} = S_{B_{\pi_i \kappa_j}}/N_{B_{\pi_i \kappa_j}}$. That is, we look at the residual structure assuming homogeneity in CTR values within each block under a binomial model. From Equation 3.2, the variance term involves an extra term $(N_k - 1)/(\gamma + 1)$, which implies that a monotonically increasing relationship between variance of the residuals and number of pulls would be indicative of extra-binomial variation. This is true as suggested by the increasing trend in Figure 6. Such exploratory methods that gauge extra-binomial variation are often used in data analysis (see [1], pp 554-558 for an example).

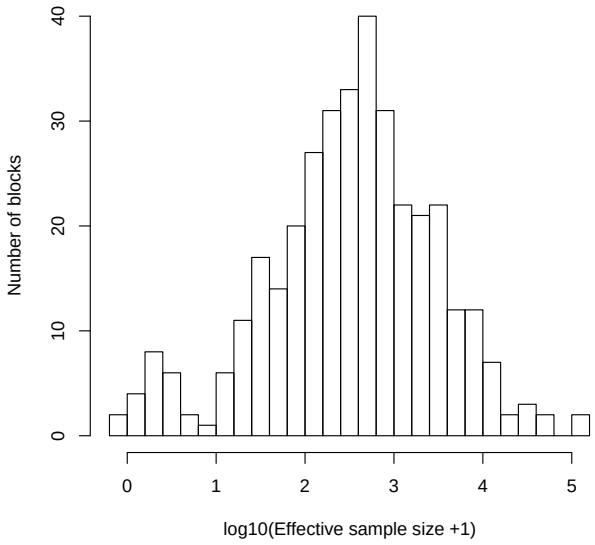


Figure 5: *Estimated effective sample sizes (γ) for blocks.*

Figure 7 shows the distribution of $\alpha_{B_{IJ}}$s (which have been linearly transformed from their true values, for reasons of data confidentiality). The relatively large values in the tail corresponds to block means for the *diagonal* blocks, i.e., blocks $B_{II}, I = 1 \ldots 20$. These contain cells for which page-class groups and ad-class groups are the same (recall that the page and ad taxonomies are identical), and so we get higher CTRs in these than in the off-diagonal blocks.

[4] we obtained a snapshot for some fixed time period

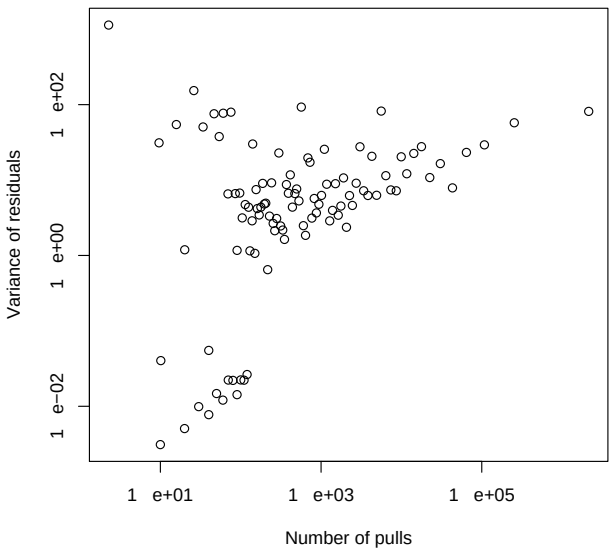


Figure 6: *Residual variance versus number of pulls:* Both the axes are on the natural logarithm scale. The variances were computed by binning number of pulls into bins of size 30

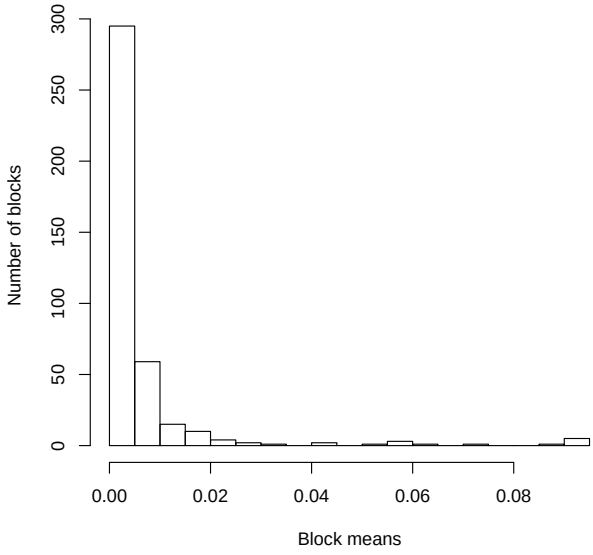


Figure 7: *Distribution of block means:* The tail values correspond to diagonal blocks.

4.3 Experiments Several CTR matrices $\mathcal{C}$ were generated using beta distributions for blocks with $\alpha_{B_{IJ}}$ values estimated as described in Section 4 and shown in Figure 7. We experimented with $\gamma = 10$ and 100. Since results were qualitatively similar, we only provide an analysis with one $\mathcal{C}$ generated with $\gamma = 10$. To facilitate extensive experimentation, we conducted our simulations on a matrix with 5×5 blocks randomly chosen from the 20 blocks reported in Section 4 resulting in a total of $50 \times 50 = 2500$ cells. To test statistical significance, we conducted 40 simulations for each policy and the length of each run was $25K$. Comparisons are made in terms of total expected revenue and total mean-squared error (MSE) that were obtained by averaging across simulations. To study the efficacy of our multi-level policy and shrinkage estimation, the following policies were considered:

- Multi-level: Our proposed policy described in Figure 2.
- *UCB1*: A policy that runs plain *UCB1* for each page class s_i. The policy is described in Figure 3.
- *Round-robin*: A policy that runs round robin for each page class s_i, i.e., for a page-class s_i, select an ad-class randomly from $R(i; +)$.
- Multi-level *w/o shrinkage*: The policy Multi-level but without using shrinkage estimation.
- *UCB1 w/ shrinkage*: We extend the *UCB1* policy to allow shrinkage estimation; i.e., the priority function can use shrinkage estimators as CTR estimates with the observed number of pulls augmented with effective sample sizes whenever a cell belongs to a block where a betabinomial model fits. Figure 8 describes the policy.

Figure 9 shows the revenue and MSE profile of *Multi-level*, *UCB1* and *Round-robin*. As expected, the proposed policy *Multi-level* does significantly better relative to the other two policies in terms of both revenue and MSE. This comprehensively proves the benefits of incorporating the taxonomy structure in the estimation and allocation processes. Our multi-level policy provides better short term revenue compared to a traditional bandit policy that treats the arms as independent.

To understand the role shrinkage plays in our procedure, Figure 10 shows revenue and MSE profiles for *Multi-level* and *Multi-level w/o shrinkage* policies. The revenue profiles for the two are almost identical. However, the MSE for *Multi-level* is significantly better than *Multi-level w/o shrinkage*. Shrinkage estimation is known to be a variance reduction technique and has proved effective in several large scale data mining applications (see [22] for an example). Hence, it is no surprise that shrinkage helps us learn CTR distributions

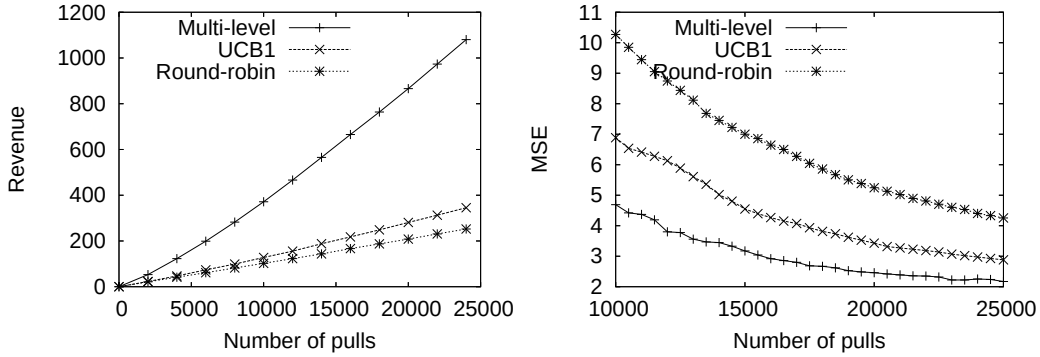
POLICY 3. (*UCB1 w/ shrinkage)*

Select the best ad-class k^* corresponding to a page-class s_i as follows:

$$k^* = argmax_{k \in R(i;+)} \left(\tilde{p}_{ik} + \sqrt{\frac{2ln(N_{R(i;+)}) + \gamma_{R(i;+)}}{N_{ik} + \gamma_{ik}}} \right).$$

Here, $\tilde{p}_{ik}$ is the posterior mean if the betabinomial model fits in $B_{\pi_i \kappa_k}$ and γ_U is the estimated effective sample size for set U. For $U = ik$, $\gamma_U = \gamma_{B_{\pi_i \kappa_k}}$ and for $U = R(i;+)$, $\gamma_U = v\gamma_{B_{\pi_i \kappa_k}}$. If the betabinomial model does not fit in $B_{\pi_i \kappa_k}$, $\tilde{p}_{ik} = \hat{p}_{ik}$ and $\gamma_U = 0$ for $U = ik, R(i;+)$.

Figure 8: *UCB1 w/ shrinkage* policy



(a) Revenue profile (b) MSE profile

Figure 9: *Total revenue and MSE versus number of pulls averaged across* 40 *simulations. All results are statistically significant, confidence intervals omitted to avoid clutter.*

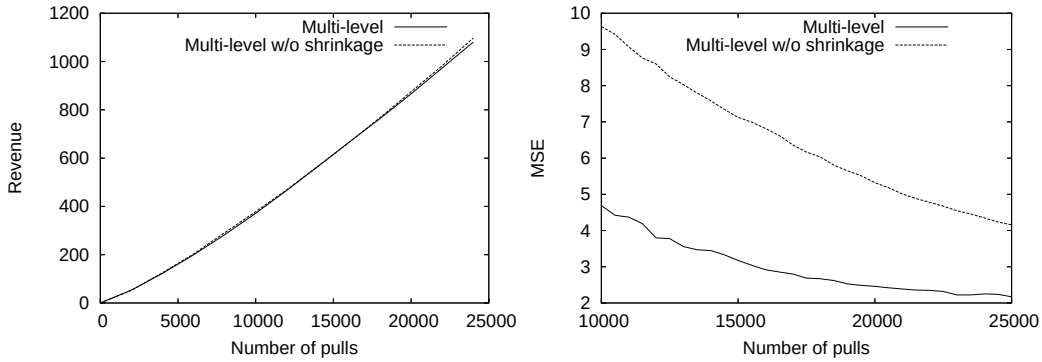
more accurately but it does so without incurring a penalty in terms of lost revenue. As emphasized in Section 1, this is extremely useful since for content match, we deal with finite inventory and hence it is beneficial to learn the CTR distribution more accurately for large number of cells. The Bayesian model when combined with our multi-stage sampling scheme provides additional accuracy with almost no extra cost.

To further understand the effect of multi-stage sampling and shrinkage, we compared policy *UCB1 w/ shrinkage* with *UCB1* which does not use shrinkage. Note that both these policies use a single stage allocation scheme. Results are shown in Figure 11. Although the MSE with shrinkage is excellent, its revenue profile deteriorates after the first $6K$ pulls. This occurs because the betabinomial model starts fitting the diagonal blocks earlier than in the more homogeneous off-diagonal blocks. Since the priority for cells belonging to a betabinomial block incorporates the effective sample size, the *UCB1 w/ shrinkage* scheme tends to explore the poor off-diagonal blocks more often than *UCB1*, thus explaining its poor revenue but better MSE. This clearly shows that shrinkage alone may not lead to better performance; it is essential to combine it with multi-stage allocation as we propose in our *Multi-level* policy for achieving good MSE without compromising too much on revenue.

5 Related Work

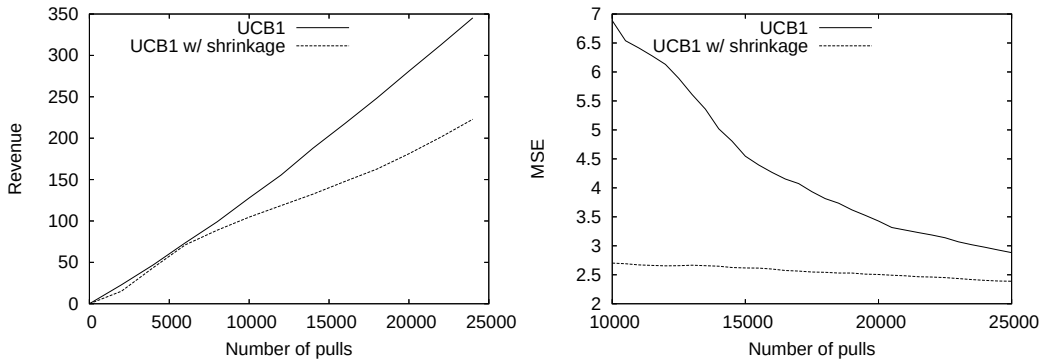
There is substantial literature on reinforcement learning [9] and multi-armed bandits [3] that is related to our problem. In our context, the classical multi-armed bandit formulation assumes independence of pages and ads resulting in a standard k-armed bandit for each page. This is a well-studied problem with a number of algorithms and heuristics whose theoretical properties are well understood [3]. However, to the best of our knowledge, generalizations for matching hierarchical feature spaces have not been studied before. Moreover, although the existing solutions are optimal asymptotically, convergence in the context of content match is slow, hence limiting their practical utility.

Another related approach are methods based on Markov Decision Processes (MDPs) [18]. Although several approaches exist for learning MDPs, the one based on Q-learning [2] is popular and well understood. In our con-



(a) Revenue profile (b) MSE profile

Figure 10: *Total revenue and MSE versus number of pulls averaged across* 40 *simulations.*



(a) Revenue profile (b) MSE profile

Figure 11: *Total revenue and MSE versus number of pulls averaged across* 40 *simulations.*

text, pages and ads correspond to the states and actions of an MDP and Q-learning selects an ad that maximizes a discounted sum of immediate reward and the optimal reward accumulated subsequently. However, the transition probability to the next page is independent of the current page and the action taken thereof, since page arrival is an independent random process that does not depend on prior arrivals and allocations. Hence, such an approach is greedy and tends to pick ads with the best current CTR estimates, potentially failing to exploit more profitable matches in the long run. We are also close in spirit to methods that learn the environment of an MDP using a model based approach to provide quicker estimates of the Q-function [17, 19]. However, the modeling framework is primarily tailored towards learning the table of state transition probabilities which does not arise in our context. Our focus is primarily on exploiting the hierarchical structure of the reward space. An alternate approach that has been pursued in the context of a large MDP is based on the sparse sampling of the look ahead tree [12].

A closely related approach are bandit solutions in the presence of covariates or side information[21] where much of the focus has been on bandit problems with small number of arms. For instance, [13] considered a one-armed bandit in a Bayesian setting; [6] extended the model of [13] to exponential family; [24] provide a policy to estimate the functional relationship between response and covariates non-parametrically for a multi-armed bandit problem. We can map our problem to this setting by regarding the groupings induced by hierarchies as covariates. However, incorporating dependencies induced by the hierarchical structure of the reward space introduce additional nuances. Moreover, the scale of our problem and data sparseness issues have not been considered in the *bandits with covariates* setting. We provide a solution to both these issues by taking recourse to a generative model for the process based on a hierarchical Bayesian framework coupled with a new multi-allocation strategy.

6 Discussion

This paper introduced a novel learning problem of matching feature spaces that are high dimensional but organized as hierarchies. An assumption that usually holds in practice is that dependencies in the reward structure are induced by lineages. We provide a bandit formulation of the problem and propose a novel policy called *Multi-level* which provides better short term revenue and accurate estimates relative to policies that do not exploit the structure in reward space. Our policy estimates the reward distribution in a Bayesian model based framework and modifies an existing k-armed bandit policy to a multi-stage allocation policy. We prove asymptotic convergence of our policy and demonstrate its superior performance through simulation experiments.

We are currently considering several extensions. We find significantly better performance with multi-stage sampling relative to single stage sampling in our simulations and hope to provide finite sample properties of our policy analytically under some simplifying assumptions. Such results are non-trivial to derive due to complex dependencies that are induced by multi-stage sampling and shrinkage estimation in the sampling path of our policy. Investigating properties of adaptive designs through simulation is a standard practice and has been extensively pursued, for instance, in the context of clinical trials; especially for designs that deal with complex issues encountered by practitioners (see [23] for an example).

We are considering a multi-stage generalization of policies other than *UCB1*. We are currently experimenting with a class of attractive policies for exponential families that are described in [7]. These policies are simple to implement and have optimal frequentist and Bayesian properties.

We have discussed a framework to match page-classes to ad-classes. In practice, we need to incorporate additional attribute information. In principle, this can be done in the betabinomial framework by introducing regression; i.e., the block means $\alpha_{B_{ij}}$s are functions of context attributes which are learned over time. However, methods to conduct this learning efficiently in a sequential manner is non-trivial. To generalize our approach to the entire taxonomy, we are pursuing a *top-down* approach where the learning starts at the coarsest level and successively switches to finer levels with increasing sample size. Other generative models (e.g., gamma-Poisson) can be easily handled by our procedure. We are also working on methods to relax the one-to-one page-class to ad-class mapping restriction.

References

[1] A.Agresti. *Categorical Data Analysis*. Wiley, New York, US, 2002.

[2] C.J.C.H.Watkins and P.Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

[3] D.A.Berry and B.Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London,UK, 1985.

[4] J.C.Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41:148–177, 1979.

[5] J.H.Albert. Computational methods using a bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, 83:1037–1044, 1988.

[6] J.Sarkar. One-armed bandit problems with covariates. *Annals of Statistics*, 19(4):1978–2002, 1991.

[7] T. L. Lai. Adaptive treatment allocation and multi-armed bandit problem. *Annals of Statistics*, 15(3):1091–1114, 1987.

[8] L.Cai and T.Hofmann. Hierarchical document categorization with support vector machines. In *ACM 13th Conference on Information and Knowledge Management*, pages 1–10, 2004.

[9] L.P.Kaelbling, M.L.Littman, and A.W.Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[10] O. Madani and D. DeCoste. Contextual recommender problems. In *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining*, pages 86–89, New York, NY, USA, 2005.

[11] M.J.Kahn and A.E.Raftery. Discharge rates of medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association*, 91:29–41, 1996.

[12] M.Kearns, Y.Mansour, and A.Y.Ng. A sparse sampling algorithm for near optimal planning in large markovian decision processes. In *Proceedings of IJCAI'99*, pages 1324–1331, 1999.

[13] M.Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74:799–806, 1979.

[14] N.Abe and A.Nakamura. Learning to optimally schedule internet banner advertisements. In *Proceedings of the 16th International Conference on Machine Learning*, 1999.

[15] N.Abe, A.W.Biermann, and P.M.Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

[16] P.Auer, N.Cesa-Bianchi, and P.Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[17] R.Dearden, N.Friedman, and D.Andre. Model based bayesian exploration. In *UAI 5th conference on Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

[18] R.E.Bellman. A markov decision process. *Journal of Mathematical Mechanics*, 6:679–684, 1975.

[19] M. Strens. A bayesian framework for reinforcement learning. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 2000.

[20] T.Lai and H.Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[21] C.-C. Wang, S. R. Kulkarni, and H. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

[22] W.DuMouchel and D.Pregibon. Empirical bayes screening for multi-item associations. In *Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[23] W.F.Rosenberger, A.N.Vidyashankar, and D.K.Agarwal. Covariate-adjusted response adaptive designs for binary response. *Journal of Biopharmaceutical Statistics*, 11:227–236, 2001.

[24] D. Yang and D.Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, 30(1):100–121, 2002.

Appendix

We begin with a set of lemmas followed by a proof in theorem 6.1.

LEMMA 6.1. *Any bandit policy which is asymptotically optimal must pull all arms infinitely often.*

Proof. First, we note that the proof is trivial for policy $UCB1$. It is impossible to stop pulling any arm j after any finite time since if we did, $log(n)/n_j \rightarrow \infty$ as $n \rightarrow \infty$. However, since the payoff probability estimate of arm j is bounded below by 0 , the priority for arm $j \rightarrow \infty$ forcing us to pull arm j by definition of $UCB1$. This leads to a contradiction. The proof for the general case follows from similar arguments but not reported here.

LEMMA 6.2. *The* Multi-level *policy will necessarily switch from using a block-level bandit in its first stage to a row-level bandit, for all page-classes that appear infinitely often and for any value of $\tau(>0)$ and $r(\geq 1)$.*

Proof. The switch occurs when $CV_{\pi_i(r)} \leq \tau$, where r and τ are parameters to the *Multi-level* policy and s_i is the page-class. Now, $CV_U = \sqrt{\frac{1-\hat{p}_U}{N_U \hat{p}_U}} \rightarrow 0$ as $N_U \rightarrow \infty$, and Lemma 6.1 implies every block $B_{\pi_i J}$ with $J \in \{1, \ldots, k_2\}$ will be pulled infinitely often, hence $N_{B_{\pi_i J}} \rightarrow \infty$. Thus, asymptotically, this switch will always occur.

LEMMA 6.3. *After switching to the row-level bandit, the observed success probability of any row-arm $R(i; B_{\pi_i J}), J = 1, \cdots, k_2$ converges to the maximum success probability among all cells in it, if the corresponding page-class s_i appears infinitely often.*

Proof. After the switch, each $R(i; B_{\pi_i J})$,$J = 1, \cdots, k_2$ is pulled infinitely often (Lemma 6.2). Let cell $(ij_{opt}) \in R(i; B_{\pi_i J})$ be the cell with the maximum success probability in $R(i; B_{\pi_i J})$, say, p_{max}. Suppose this particular row-arm has been pulled N times in the first stage of the row-level bandit. Each time, our cell-level policy is run in the second stage; this is exactly *UCB1* [16] but with a constant $\gamma_{\pi_i J}$ "prior" observations for each cell inside this row-arm. Since *UCB1* has $O(logN)$ regret, cell (ij_{opt}) is pulled at least $N - c \cdot \log N$ times, for some constant c. Thus, the observed success probability of this row-arm is $p_{row} \geq p_{max} \cdot \frac{N - c \cdot \log N}{N} \rightarrow p_{max}$, since $N \rightarrow \infty$ by Lemma 6.2.

THEOREM 6.1. *Policy* Multi-level *is asymptotically optimal for each $\tau(>0)$ and $r(\geq 1)$.*

Proof. From Lemma 6.2, the switch to row-level will happen, and from Lemma 6.3, the best row-arm (row containing the best cell) and the best cell itself are correctly identified asymptotically. The result follows.

In practice, early switching to row-level may lead to poor estimation of block priors, so a conservative switch is used to ensure accurate estimation.