Exploration Scavenging

John Langford JL@YAHOO-INC.COM
Alexander Strehl STREHL@YAHOO-INC.COM
Yahoo! Research, 111 W. 40th Street, New York, New York 10018

Jennifer Wortman WORTMANJ@SEAS.UPENN.EDU
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

Abstract

We examine the problem of evaluating a policy in the contextual bandit setting using only observations collected during the execution of another policy. We show that policy evaluation can be impossible if the exploration policy chooses actions based on the side information provided at each time step. We then propose and prove the correctness of a principled method for policy evaluation which works when this is not the case, even when the exploration policy is deterministic, as long as each action is explored sufficiently often. We apply this general technique to the problem of offline evaluation of internet advertising policies. Although our theoretical results hold only when the exploration policy chooses ads independent of side information, an assumption that is typically violated by commercial systems, we show how clever uses of the theory provide non-trivial and realistic applications. We also provide an empirical demonstration of the effectiveness of our techniques on real ad placement data.

1. Introduction

The k-armed bandit problem (Lai & Robbins, 1985; Berry & Fristedt, 1985; Auer et al., 2002; Even-Dar et al., 2006) has been studied in great detail, primarily because it can be viewed as a minimal formalization of the exploration problem faced by any autonomous agent. Unfortunately, while its minimalism admits tractability and insight, it misses details that are necessary for application to many realistic problems. For instance, the problem of internet advertising can be viewed as a type of bandit problem in which choosing an ad or set of ads to display corresponds to choosing an arm to pull. However, this formalization is inadequate in practice, as vital information is ignored. In particular, a successful ad placement policy might choose ads based on the content of the web page on which the ads are displayed. The standard k-armed bandit formulation ignores this useful information.

This shortcoming can be rectified by modeling the problem as an instance of the *contextual bandit problem* (Langford & Zhang, 2007), a generalization of the k-armed bandit problem that allows an agent to first observe an *input* or *side information* before choosing an arm. This problem has been studied under different names, including associative reinforcement learning (Kaelbling, 1994), bandits with side information (Wang et al., 2005), and bandits with experts (Auer et al., 1995), yet its analysis is far from complete.

In this paper, we study *policy evaluation* in the contextual bandit setting. Policy evaluation is the problem of evaluating a new strategy for behavior, or *policy*, using only observations collected during the execution of another policy. The difficulty of this problem stems from the lack of control over available data. Given complete freedom, an algorithm could evaluate a policy simply by executing it for a sufficient number of trials. However, in real-world applications, we often do not have the luxury of executing arbitrary policies, or we may want to distinguish or search among many more policies than we could evaluate independently.

We begin by providing impossibility results characterizing situations in which policy evaluation is *not* possible. In particular, we show that policy evaluation can be impossible when the exploration policy depends on the current input. We then provide and prove the correctness of a principled method for policy evaluation when this is not the case. This technique, which we

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

call "exploration scavenging," can be used to accurately estimate the value of any new policy as long as the exploration policy does not depend on the current input and chooses each action sufficiently often, even if the exploration policy is deterministic. The ability to depend on deterministic policies makes this approach more applicable than previous techniques based upon known and controlled randomization in the exploring policy. We also show that exploration scavenging can be applied if we wish to choose between multiple policies, even when these policies depend on the input, which is a property shared by most real ad-serving policies. This trick allows exploration scavenging to be applied to a broader set of real-life problems.

The motivating application for our work is *internet advertising*. Each time a user visits a web page, an advertising engine places a limited number of ads in a *slate* on the page. The ad company receives a payment for every ad clicked by the user. Exploration scavenging is well-suited for this application for a few reasons. First, an advertising company may want to evaluate a new method for placing ads without incurring the risk and cost of actually using the new method. Second, there exist logs containing huge amounts of historical click data resulting from the execution of existing ad-serving policies. It is economically sensible to use this data, if possible, when evaluating new policies.

In Section 4, we discuss the application of our methods to the ad display problem, and present empirical results on data provided by Yahoo!, a web search company. Although this application actually violates the requirement that the exploration policy be independent of the current input, the techniques show promise, leading us to believe that exploration scavenging can be useful in practice even when the strong assumptions necessary for the theoretical results do not hold.

To our knowledge, the only similar application work that has been published is that of Dupret et al. (2007) who tackle a similar problem from a Bayesian perspective using different assumptions which lead to different solution techniques. Our approach has the advantage that the estimated value is the output of a simple function rather than an EM optimization, facilitating interpretation of the evaluation method.

2. The Contextual Bandit Setting

Let $\mathcal{X}$ be an arbitrary input space, and $\mathcal{A} = \{1, \cdots, k\}$ be a set of actions. An instance of the contextual bandit problem is specified by a distribution D over tuples $(x, \vec{r})$ where $x \in \mathcal{X}$ is an input and $\vec{r} \in [0,1]^k$ is a vector of rewards. Events occur on a round by round basis where on each round t:

1. The world draws $(x_t, \vec{r}_t) \sim D$ and announces x_t.
2. The algorithm chooses an action $a_t \in \mathcal{A}$, possibly as a function of x_t and historical information.
3. The world announces the reward r_{t,a_t} of action a_t.

The algorithm does not learn what reward it would have received if it had chosen an action $a \neq a_t$.

The standard goal in this setting is to maximize the sum of rewards r_{t,a_t} over the rounds of interaction. An important subgoal, which is the focus of this paper, is policy evaluation. Here, we assume that we are given a data set $S \in (\mathcal{X} \times \mathcal{A} \times [0,1])^T$, which is generated by following some fixed policy π for T steps. Now, given a different policy $h : \mathcal{X} \to \mathcal{A}$, we would like to estimate the *value* of policy h, that is,

$$V_D(h) := E_{(x,\vec{r})\sim D}[r_{h(x)}].$$

The standard k-armed bandit is a special case of the contextual bandit setting in which $|\mathcal{X}| = 1$.

3. Evaluating Policies

In this section, we characterize situations in which policy evaluation may not be possible, and provide techniques for estimating the value of a policy when it is. To start, we show that when the exploration policy π depends on the input x, policy evaluation can be impossible. Later, we show that when the exploration policy π has no dependence on the current input, there exists a technique for accurately estimating the value of a new policy h as long as the exploration policy chooses each action sufficiently often. Finally, we show that exploration scavenging can be applied in the situation in which we are choosing between multiple exploration policies, even when the exploration policies themselves depend on the current input.

3.1. Impossibility Results

First, note that policy evaluation is not possible when the exploration policy π chooses some action a with zero probability. This is true even in the standard k-armed bandit setting. If the exploration policy always chooses action 1, and the policy to evaluate always chooses action 2, then policy evaluation is hopeless.

It is natural to ask if it is possible to build a policy evaluation procedure that is guaranteed to accurately evaluate a new policy given data collected using an arbitrary exploration policy π as long as π chooses each action sufficiently often. The following theorem shows that this goal is unachievable. In particular, it shows

that if the exploration policy π depends on the current input, then there are cases in which new policies h cannot be evaluated using observations gathered under π, even if π chooses each action frequently. Specifically, there can exist two contextual bandit distributions D and D' that result in indistinguishable observation sequences even though $V_D(h)$ and $V_{D'}(h)$ are far apart. Later we show that in the same context, if we disallow input-dependent exploration policies, policy evaluation becomes possible

Theorem 1 *There exist contextual bandit problems D and D' with $k = 2$ actions, a hypothesis h, and a policy π dependent on the current observation x_t with each action visited with probability $1/2$, such that observations of π on D are statistically indistinguishable from observations of π on D', yet $|V_D(h) - V_{D'}(h)| = 1$.*

Proof: The proof is by construction. Suppose x_t takes on the values 0 and 1, each with probability 0.5 under both D and D'. Let $\pi(x) = x$ be the exploration policy, and let $h(x) = 1 - x$ be the policy we wish to evaluate. Suppose that rewards are deterministic given x_t, as summarized in the following table.

	Under D		Under D'	
	$r_{t,0}$	$r_{t,1}$	$r_{t,0}$	$r_{t,1}$
$x_t = 0$	0	0	0	1
$x_t = 1$	0	1	1	1

Then $V_D(h) = 0$, while $V_{D'}(h) = 1$, but observations collected using exploration policy π are indistinguishable for D and D'. ■

3.2. Techniques for Policy Evaluation

We have seen that policy evaluation can be impossible in general if the exploration policy π depends on the current input or fails to choose each action sufficiently often. We now discuss techniques for policy evaluation when this is not the case. Theorem 2 shows that in some very special circumstances, it is possible to create an unbiased estimator for the value of a policy h using exploration data from another policy. The main result of this section, Theorem 3, shows that this estimator is often close to the value of the policy, even when the stringent conditions in the Theorem 2 are not satisfied.

Theorem 2 *For any contextual bandit distribution D over $(x, \vec{r})$, any policy h, any exploration policy π such that (1) for each action a, there is a constant $T_a > 0$ for which $|\{t : a_t = a\}| = T_a$ with probability 1, and (2) π chooses a_t independent of x_t,*

$$V_D(h) = E_{\{x_t,\vec{r}_t\}\sim D^T}\left[\sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right].$$

Proof:

$$\begin{aligned}
& E_{\{x_t,\vec{r}_t\}\sim D^T}\left[\sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right] \\
&= E_{\{x_t,\vec{r}_t\}\sim D^T}\left[\sum_{a=1}^{k} \sum_{\{t:a_t=a\}} \frac{r_{t,a} I(h(x_t) = a)}{T_a}\right] \\
&= \sum_{a=1}^{k} E_{\{x_t,\vec{r}_t\}\sim D^T}\left[\sum_{\{t:a_t=a\}} \frac{r_{t,a} I(h(x_t) = a)}{T_a}\right] \\
&= \sum_{a=1}^{k} E_{(x,\vec{r})\sim D}\left[T_a \frac{r_a I(h(x) = a)}{T_a}\right] \\
&= E_{x,\vec{r}\sim D}\left[\sum_{a=1}^{k} r_a I(h(x) = a)\right] = V_D(h).
\end{aligned}$$

The first equality is a reordering of the sum. The second and fourth follow from linearity of expectation.

The third equality is more subtle. Consider a fixed action a. The term $\sum_{\{t:a_t=a\}} r_{t,a} I(h(x_t) = a)/T_a$ involves drawing T bandit samples $(x_t, \vec{r}_t)$ and summing the term $r_{t,a} I(h(x_t) = a)/T_a$ over only the times t for which the exploration policy chose action a. There are precisely T_a such trials. The equality then follows from the fact that the quantity $r_{t,a} I(h(x_t) = a)/T_a$ is identically distributed for all t such that $a_t = a$. It is critical that the exploration policy π chooses a_t independent of x_t (to make the numerator identical) and that T_a is fixed (to make the denominator identical). If a_t depends on x_t, then these values are no longer identically distributed and the equality does not hold. This is important, as we have seen that evaluation is not possible in general if a_t can depend on x_t. ■

Conditions (1) and (2) in the theorem are satisfied, for example, by any policy which visits each action and chooses actions independent of observations. This theorem represents the limit of what we know how to achieve with a strict equality. It can replace the sample selection bias (Heckman, 1979) lemma used in the analysis of the Epoch-Greedy algorithm (Langford & Zhang, 2007), but cannot replace the analysis used in EXP4 (Auer et al., 1995) without weakening their theorem statement to hold only in IID settings.

The next theorem, which is the main theoretical result of this paper, shows that in a much broader set of circumstances, the estimator in the previous lemma is useful for estimating $V_D(h)$. Specifically, as long as the exploration does not depend on the current input and chooses each action sufficiently frequently, the estimator can be used for policy evaluation.

Theorem 3 *For every contextual bandits distribution D over $(x, \vec{r})$ with rewards $r_a \in [0,1]$, for every sequence of T actions a_t chosen by an exploration policy π that may be a function of history but does not depend on x_t, for every hypothesis h, then for any $\delta \in (0,1)$, with probability $1-\delta$,*

$$\left| V_D(h) - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}} \right| \leq \sum_{a=1}^{k} \sqrt{\frac{2\ln(2kT/\delta)}{T_a}}$$

where $T_a = |\{t : a_t = a\}|$.

Proof: First notice that

$$V_D(h) = \sum_{a=1}^{k} E_{x,\vec{r}\sim D}\left[r_a I(h(x) = a)\right]. \qquad (1)$$

Fix an action a. Let t_i denote the ith time step that action a was chosen, with $t_i = 0$ if $i > T_a$. Note that t_i is a random variable. For $i = 1, \ldots, T$ define

$$Z_i = \begin{cases} r_{t_i,a} I(h(x_{t_i}) = a) \\ \qquad -E_{x,\vec{r}\sim D}[r_a I(h(x) = a)] & \text{if } i \leq T_a, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $Z_i \in [-1,1]$ and $E[Z_i] = 0$ for all i. Now fix a positive integer $t \in \{1, \ldots, T\}$. We apply Azuma's inequality (see, for example, Alon and Spencer (2000)) to show that for any $\delta' \in (0,1)$, with probability $1-\delta'$,

$$\frac{1}{t}\left|\sum_{i=1}^{t} Z_i\right| \leq \sqrt{\frac{2\ln(2/\delta')}{t}}\ , \qquad (2)$$

and so if $t \leq T_a$,

$$\left| E_{x,\vec{r}\sim D}[r_a I(h(x) = a)] - \frac{1}{t}\sum_{i=1}^{t} r_{t_i,a} I(h(x_{t_i}) = a) \right|$$

is upper bounded by $\sqrt{2\ln(2/\delta')/t}$. Applying the union bound with $\delta' = \delta/(Tk)$, we see that Equation 2 holds for *all* $t \in \{1, \ldots, T\}$ and thus for $t = T_a$ with probability δ/k. Applying the union bound again yields a bound that holds for all actions. Summing over actions and applying Equation 1 yields the lemma. ∎

Note that the counter-example given in Theorem 1 satisfies all conditions of Theorem 3 except for the assumption on π. Thus, we cannot solve the policy exploration problem, in general, unless we make assumptions that limit π's dependence on input.

Corollary 4 *For every contextual bandit distribution D over $(x, \vec{r})$, for every exploration policy π choosing action a_t independent of the current input, for every policy h, if every action $a \in \{1, \cdots, k\}$ is guaranteed to be chosen by π at least a constant fraction of the time, then as $T \to \infty$, the estimator*

$$\hat{V}_D(h) = \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}$$

grows arbitrarily close to $V_D(h)$ with probability 1.

These observations can be utilized in practice in a simple way. Given a data set S of observations (x_t, a_t, r_{a_t}) for $t = \{1, \cdots, T\}$, we can calculate $\hat{V}_D(h)$ as above and use this as an estimator of $V_D(h)$. For sufficiently large data sets S, as long as each action is chosen sufficiently often, this estimator is accurate.

3.3. Tighter Bounds in Special Cases

In some special cases when there is sufficient randomization in the exploration policy or the policy h, it is possible to achieve tighter bounds using a slightly modified estimator. Theorem 5 shows that the dependence on the number of actions k can be improved in the special case in which $\Pr(h(x) = a_t) = 1/k$ independent of x. This is true, for instance, when either the exploration policy π or the policy h chooses actions uniformly at random. We suspect that tighter bounds can be achieved in other special cases as well.

Theorem 5 *For every contextual bandits distribution D over $x, \vec{r}$ with rewards $r_a \in [0,1]$, for every sequence of actions a_t chosen by an exploration policy π that may be a function of history but does not depend on x_t and every hypothesis h, if $\Pr(h(x) = a_t) = 1/k$ independent of x and if $|\{t : a_t = a\}| > 0$ for all a, then for any $\delta \in (0,1)$, with probability $1-\delta$,*

$$\left| V_D(h) - \sum_{t=1}^{T} \frac{k r_{t,a_t} I(h(x_t) = a_t)}{T} \right| \leq k\sqrt{\frac{2\ln(2k/\delta)}{T}}\ .$$

Proof: Since we have assumed that $\Pr(h(x_t) = a_t) = 1/k$ independent of x_t,

$$\begin{aligned} V_D(h) &= E_{x,\vec{r}\sim D}[k r_{h(x)} I(h(x) = a_t)] \\ &= E_{x,\vec{r}\sim D}[k r_{a_t} I(h(x) = a_t)]\ . \end{aligned}$$

For all t, define

$$Z_t = k r_{t,a_t} I(h(x_t) = a_t) - E_{x,\vec{r}\sim D}[k r_{a_t} I(h(x) = a_t)]\ .$$

$Z_t \in [-k, k]$ and $E[Z_{t,a}] = 0$. Applying Azuma's inequality and the union bound yields the lemma. ∎

3.4. Multiple Exploration Policies

So far, all of our positive theoretical results have required that the exploration policy choose actions independent of the current input. There do exist special cases in which exploration data is provably useful for policy evaluation even when the exploration policy depends on context. We briefly describe one such case.

Suppose we have collected data from a system that has rotated through K known exploration policies $\pi_1, \pi_2, \cdots, \pi_K$ over time. For example, we may have logs of historical ad display data from a company that has used K different ad display policies. Each individual exploration policy π_i may depend on context, but we assume that the choice of which policy was used by the system at any given time does not.

We can redefine the action of the bandit problem as a choice of one of the K base policies to follow; action a_i now corresponds to choosing the ad chosen by policy π_i. Since historically the decision about which policy to use was made independent of the context x, we can view the exploration policy as oblivious with respect to x. Theorem 3 then implies that we can accurately estimate the value of any policy π which chooses from the set of actions chosen by the K base policies. This can be more powerful than competing with each historic policy, because π can make context-dependent choices about which policy to follow, potentially achieving better performance than any single policy.

4. Application to Internet Advertising

Technology companies are interested in finding better ways to search both over the myriad pages of the internet and over the increasingly large selection of potential ads to display. However, given a candidate algorithm (or *ad-serving policy* in the case of online advertising), a company faces a real-life "exploration-exploitation" dilemma. The new algorithm could be better than existing ones, but it could be worse. To evaluate the performance of an algorithm, the company might decide to adopt it for a short time on a subset of web traffic. This method produces accurate estimates of performance, but the evaluation phase can be costly in terms of lost revenue if the candidate algorithm performs poorly, and this cost grows linearly in the number of candidate algorithms that the company would like to evaluate. Clearly, a method of determining the strengths or weaknesses of an algorithm without adopting it would be highly useful.

In this section, we tackle the problem of evaluating a new ad-serving policy using data logged from an existing system. We state our results in terms of the online advertising problem, but everything we discuss can be applied to web search with little or no modification.

We begin by showing how to directly apply exploration scavenging techniques to the problem, and discuss the primary drawbacks of this simple approach. Instead, we consider a standard simplifying assumption whose adoption leads to a more realistic method for policy evaluation. This assumption, that click-through rates are factorable, leads to another interesting theoretical problem, estimating the attention decay coefficients of the click-through rates, which can also be accomplished using techniques from Section 3.

4.1. The Direct Approach

The online advertising problem can be directly mapped to the contextual bandit problem, allowing us to apply results from Section 3. Here the input space is the universe of all possible web pages and the action space contains all slates of ads. The reward is a bit vector that identifies whether or not each returned ad was clicked.[1] This bit vector can be converted to a single real-valued reward r in a number of ways, for instance, by simply summing the components, yielding the total number of clicks received, and normalizing. The example would then be used to compute a number $r \cdot I(h(x) = s)/\text{Count}(s)$, where $\text{Count}(s)$ is the number of times the slate s was displayed during all trials. According to Theorem 3, summing this quantity over all trials yields a good estimator of the value of the new policy h.

There is a significant drawback to this approach. Due to the indicator variable, the contribution to the sum for a single example is zero unless $h(x) = s$, which means that the slate chosen by the candidate algorithm h is *exactly* the same as the slate produced by the current system π. With a large set of ads and a large slate size, it is very unlikely that the same slate is chosen many times, and thus the resulting estimator for the value of h has an extremely high variance and may not exist for most slates. In the next section, we show how a standard assumption in the online advertising community can be used to reduce the variance.

4.2. The Factoring Assumption

The problem described above can be avoided by making a *factoring assumption*. Specifically, we assume that the probability of clicking an ad can be decomposed or factored into two terms, an intrinsic *click-through rate* (CTR) that depends only on the web

[1] The reward function can be modified easily to reflect the actual revenue generated by each click.

page x and the ad a, and a position-dependent multiplier C_i for position i, called the *attention decay coefficient* (ADC). This assumption is commonly used in the sponsored search literature (Borgs et al., 2007; Lahaie & Pennock, 2007).

Formally, let $\mathcal{P}(x, a, i)$ be the probability that ad a is clicked when placed in position i on web page x. We assume that $\mathcal{P}(x, a, i) = C_i \cdot \mathcal{P}(x, a)$, where $\mathcal{P}(x, a)$ is the intrinsic (position independent) click-through rate for ad a given input x, and C_i is a position-dependent constant. Here $C_1 = 1$, so $\mathcal{P}(x, a) = \mathcal{P}(x, a, 1)$.

A key observation is that this assumption allows us to transition from dealing directly with slates of ads to focusing on single ads. Let ℓ be the number of ads shown in a slate. Given an example $(x, s, \vec{r})$, we can form ℓ transformed examples of the form (x, a_i, r'_i) where a_i is the ith ad in the slate and $r'_i = r_i/C_i$. In other words, r'_i is $1/C_i$ if the ith ad was clicked, and 0 otherwise; the division by the ADC puts the rewards on the same scale, so the expected value of the reward for a fixed pair (x, a_i) is $\mathcal{P}(x, a_i)$.

Let $\sigma(a, x)$ be the slot in which the evaluation policy h places ad a on input x; if h does not display a on input x, then $\sigma(a, x) = 0$. For convenience, define $C_0 = 0$. We define a new estimator of the value of h as

$$\hat{V}_D(h) = \sum_{t=1}^{T} \sum_{i=1}^{\ell} \frac{r'_i C_{\sigma(a_i,x)}}{T_{a_i}} \ , \tag{3}$$

where T_a is the total number of impressions received by a (i.e., the total number of times add a is displayed). Here C_i takes the place of the indicator function used in the estimates in Section 3, giving higher weights to the rewards of ads that h places in better slots.

Using the results from Section 3, it is straight-forward to show that this estimator is consistent as long as the current ad-serving policy does not depend on the input webpage x and every ad is displayed. However, to apply this transformation, we require knowledge of the ADCs. In the next section we show how to estimate them, again using nonrandom exploration.

4.3. Estimating Attention Decay Coefficients

Assume that a data set S is available from the execution of an ad-serving policy π that chooses the tth slate of ads to display independent of the input x_t (though possibly dependent on history). As before, S includes observations $(x_t, \vec{a}_t, \vec{r}_{t,a_t})$ for $t = \{1, \cdots, T\}$, where $\vec{a}_t$ is the slate of ads displayed at time t and $\vec{r}_{t,a_t}$ is the reward vector. Our goal is to use this data to estimate the attention decay coefficients $C_2, \ldots, C_\ell$.

We first discuss a naive ADC estimator, and then go on to show how it can be improved. In the following sections, let $C(a, i)$ be the number of clicks on ad a observed during rounds in which ad a is displayed in position i. Let $M(a, i)$ be the number of impressions of ad a in slot i, i.e., the number of times that the exploration policy chooses to place ad a in slot i. Finally, let $\mathrm{CTR}(a, i) = C(a, i)/M(a, i)$ be the observed click-through rate of ad a in slot i, with $\mathrm{CTR}(a, i)$ defined to be 0 when $M(a, i) = 0$.

4.3.1. The Naive Estimator

Initially, one might think that the ADCs can be calculated by taking the ratio between the global empirical click-through rate for each position i and the global empirical click-through rate for position 1. Formally,

$$\mathrm{Est}_{\mathrm{naive}}(\mathrm{i}) := \frac{\sum_a C(a, i) / \sum_a M(a, i)}{\sum_a C(a, 1) / \sum_a M(a, 1)}.$$

Unfortunately, as we will see in Section 4.4, this method has a bias which is often quite large in practice. In particular, it often underestimates the ratios C_i due to the fact that existing ad-serving policies generally already place better ads (with higher $\mathcal{P}(x, a)$) in the better slots. To overcome this bias, we must design a new estimator.

4.3.2. A New Estimator

Consider a fixed ad a and a fixed position $i > 1$. Clearly if a is placed in position i sufficiently many times, it is possible to estimate the probability of a being clicked in position i fairly accurately. If we also estimate the corresponding click-through rate for ad a in position 1, we may estimate C_i using a ratio of these two click-through rates, since $C_i = E_{x\sim D}[\mathcal{P}(x, a, i)]/E_{x\sim D}[\mathcal{P}(x, a, 1)]$. If we perform this procedure for all ads, we can average the resulting estimates to form a single, typically very accurate, estimate. Formally, we propose an estimator of the form

$$\mathrm{Est}_{\vec{\alpha}}(i) = \frac{\sum_a \alpha_a \mathrm{CTR}(a, i)}{\sum_a \alpha_a \mathrm{CTR}(a, 1)}, \tag{4}$$

where $\vec{\alpha}$ is a vector of nonnegative constants α_a for each ad $a \in \mathcal{A}$.

Theorem 6 *If the ad-display policy chooses slates independent of input and $\vec{\alpha}$ has all positive entries, then the estimator $\mathrm{Est}_{\vec{\alpha}}$ in Equation 4 is consistent.*

Proof: Consider any fixed ad a and position i, and suppose that we are only interested in revenue generated by position i. Let h be the constant hypothesis that always places ad a in position i. $V_D(h)$ is then

$E_{x\sim D}\mathcal{P}(x,a,i)$. From Corollary 4, it is clear that

$$\hat{V}_D(h) = \sum_{t=1}^{T} \frac{(r_t I(h(x_t) = a_t))}{|\{t' : a_{t'} = a_t\}|}$$

converges to $V_D(h)$. Here $\hat{V}_D(h)$ is precisely $\mathrm{CTR}(a,i)$, so $\mathrm{CTR}(a,i)$ converges to $E_{x\sim D}\mathcal{P}(x,a,i)$ for all a and i. This implies that $\mathrm{Est}_{\vec{\alpha}}(p)$ converges to

$$\frac{\sum_a \alpha_a E_{x\sim D}\mathcal{P}(x,a,i)}{\sum_a \alpha_a E_{x\sim D}\mathcal{P}(x,a,1)} = \frac{\sum_a \alpha_a E_{x\sim D} C_i \mathcal{P}(x,a)}{\sum_a \alpha_a E_{x\sim D} C_1 \mathcal{P}(x,a)} = \frac{C_i}{C_1} = C_i \ .$$

■

Theorem 6 leaves open the question of how to choose $\vec{\alpha}$. If every component of $\vec{\alpha}$ is set to the same value, then the estimate for C_i can be viewed as the mean of all estimates of C_i for each ad a. However, it may be the case that the estimates for certain ads are more accurate than others, in which case we'd like to weight those more heavily. In particular, we may want to pick $\vec{\alpha}$ to minimize the variance of our final estimator. Since it is difficult to analytically compute the variance of a quotient, we approximate it by the variance of the sum of the numerator and denominator, as this tends to reduce the variance of the quotient. The proof of the following theorem is omitted due to lack of space.

Theorem 7 *The variance of the expression*

$$\sum_a \alpha_a \mathrm{CTR}(a,i) + \sum_a \alpha_a \mathrm{CTR}(a,1)$$

subject to $\sum_a \alpha_a = 1$ *is minimized when*

$$\alpha_a := \frac{2M(a,i)\cdot M(a,1)}{M(a,i)\sigma^2_{a,1} + M(a,1)\sigma^2_{a,i}},$$

where $\sigma^2_{a,i}$ *is the variance of the indicator random variable that is 1 when ad* a *is clicked given that ad* a *is placed in position* i*.*

It is undesirable that π is required to have no dependence on the current web page x_t when choosing the slate of ads to display, since most current ad-serving algorithms violate this assumption. However, as we have seen in Section 3.1, when this assumption is violated, exploration scavenging is no longer guaranteed to work. In the worst case, we cannot trust our estimated ADCs from data generated by an x-dependent π. Luckily, in practice, it is generally not the case that extreme scenarios like the counterexample in the proof of Theorem 1 arise. It is more likely that the existing ad-serving algorithm and the new algorithm choose among the same small set of ads to display for any given context (for example, the set of ads for which advertisers have placed bids for the current search term in the sponsored search setting) and the primary difference between policies is the order in which these ads are displayed. In such settings it is also the case that additional opportunities for exploration arise naturally. For example, sometimes ads run out of budget, removing them from consideration and forcing the ad-serving algorithm to display an alternate slate of ads.

4.4. Empirical comparison

We are interested in comparing the methods developed in this work to standard methods used in practice. A common technique for estimating ADCs borrowed from the information retrieval literature is discounted cumulative gain (Järvelin & Kekäläinen, 2002). In relation to our work, discounted cumulative gain (DCG) can be viewed as a particular way to specify the ADCs that is *not* data-dependent. In particular, given a parameter b, DCG would suggest defining $C_i = 1/log_b(b+i)$ for all i. As shown next, when we estimated the ADCs using our new method on a large set of data we get values that are very close to those calculated using DCG with $b = 2$.

We present coefficients that were computed from training on about 20 million examples obtained from the logs of "Content Match", Yahoo!'s online advertisement engine. Since we don't know the true variances $\sigma^2_{a,p}$ for the distributions over clicks, we heuristically assume they are all equal and use the estimator defined by $\alpha_a = M(a,p)\cdot M(a,1)/(M(a,p)+M(a,1))$. The following table summarizes the coefficients computed for the first four slots using the naive estimator and the new estimator, along with the DCG coefficients. As suspected, the coefficients for the new estimator are larger than the old, suggesting a reduction in bias.

	C_1	C_2	C_3	C_4
Naive	1.0	0.512090	0.369638	0.271847
New	1.0	0.613387	0.527310	0.432521
DCG	1.0	0.630930	0.5	0.430677

4.5. Toward A Realistic Application

To reduce the high variance of the direct application of exploration scavenging to internet advertising, we made use of the factoring assumption and derived the estimator given in Equation 3. Unfortunately this new estimator may still have an unacceptably large variance. By examining Equation 3, we observe that the method only benefits from examples in which the exploration policy and the new policy h choose overlapping sets of ads to display. When ads are drawn from

a large database, this may be too rare of an event.

Instead of considering policies which rank from the set of all ads, we can consider policies h_π reordering the ads which π chooses to display. A good reordering policy plausibly provides useful information to guide the choice of a new ranking policy.

We define an alternate estimator

$$\hat{V}_D(h_\pi) = \sum_{t=1}^{T} \sum_{i=1}^{\ell} r'_i C_{\sigma'(a_i,x)} \ ,$$

where $\sigma'(a_i, x)$ is the slot that h_π would assign to ad a_i in this new model. This method gives us an (unnormalized) estimate of the value of first using π to choose k ads to display in a slate and then using h_π to reorder them. This approach has small variance and quickly converges.

To illustrate our method we used a training set of 20 million examples gathered using Yahoo!'s current ad-serving algorithm π. We let the policy h_π be the policy that reorders ads to display those with the highest empirical click-through rate first, ignoring the context x. We used $r = C_{j'}/C_i$, (with coefficients given by the new unbiased method) to compute the number of clicks we expect the new policy (using h_π to reorder π's slate) to receive per click of the old policy π. Here j' is the relative position of ad a_i when the ads in the slate shown by π are reordered (in descending order) by h_π. This number, which was computed using a test set of about two million examples, turned out to be 1.086. When we computed the same quantity for the policy h'_π that reorders ads at random, we obtained 1.016. Thus, exploration scavenging strongly suggests using policy h_π over h'_π, matching our intuition.

5. Conclusion

We study the process of "exploration scavenging," reusing information from one policy to evaluate a new policy, and provide procedures that work *without* randomized exploration, as is commonly required. This new ability opens up the possibility of using machine learning techniques in new domains which were previously inaccessible.

Using the derandomized exploration techniques described here, we show how to estimate the value of a policy reordering displayed ads on logged data *without* any information about random choices made in the past. There are several caveats to this approach, but the results appear to be quite reasonable.

Note that this methodology is simply *impossible* without considering methods for derandomized exploration, so the techniques discussed here open up new possibilities for solving problem.

Acknowledgments

We are grateful to Michael Kearns for a useful discussion of the theoretical results, and to our anonymous reviewers for their thought-provoking questions.

References

Alon, N., & Spencer, J. (2000). *The probabilistic method*. Interscience Series in Discrete Mathematics and Optimization. John Wiley. Second edition.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite time analysis of the multi-armed bandit problem. *Machine Learning*, *47*, 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *36th Annual IEEE Symposium on Foundations of Computer Science*.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments*. London, UK: Chapman and Hall.

Borgs, C., Chayes, J., Etesami, O., Immorlica, N., Jain, K., & Mahdian, M. (2007). Dynamics of bid optimization in online advertisement auctions. *16th International World Wide Web Conference*.

Dupret, G., Murdock, V., & Piwowarski, B. (2007). Web search engine evaluation using clickthrough data and a user model. *16th Intl. World Wide Web Conference*.

Even-Dar, E., Mannor, S., & Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, *7*, 1079–1105.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*, 422–446.

Kaelbling, L. P. (1994). Associative reinforcement learning: A generate and test algorithm. *Machine Learning*, *15*.

Lahaie, S., & Pennock, D. (2007). Revenue analysis of a family of ranking rules for keyword auctions. *8th ACM Conference on Electronic Commerce*.

Lai, T., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*, 4–22.

Langford, J., & Zhang, T. (2007). The epoch greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*.

Wang, C.-C., Kulkarni, S. R., & Poor, H. V. (2005). Bandit problems with side observations. *IEEE Transactions on Automatic Control*, *50*, 338–355.