Machine Learning in an Auction Environment

Patrick Hummel
Google Inc.
Mountain View, CA, USA
phummel@google.com

R. Preston McAfee
Google Inc.
Mountain View, CA, USA
mcafeer@google.com

ABSTRACT

We consider a model of repeated online auctions in which an ad with an uncertain click-through rate faces a random distribution of competing bids in each auction and there is discounting of payoffs. We formulate the optimal solution to this explore/exploit problem as a dynamic programming problem and show that efficiency is maximized by making a bid for each advertiser equal to the advertiser's expected value for the advertising opportunity plus a term proportional to the variance in this value divided by the number of impressions the advertiser has received thus far. We then use this result to illustrate that the value of incorporating active exploration into a machine learning system in an auction environment is exceedingly small.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Algorithms, Economics, Theory

Keywords

Auctions, Explore/Exploit, Machine Learning, Online Advertising

1. INTRODUCTION

In standard Internet auctions in which bidders bid by specifying how much they are willing to pay per click, it is standard to rank the advertisers by a product of their bid and their click-through rate, or their expected cost-per-1000-impressions (eCPM) bids. While this is a sensible approach to take for determining the best ad to show for a particular query, it is potentially a suboptimal approach if one cares about showing the best possible ads in the long run. In online auctions, new ads that may compete in the auctions are constantly entering the system, and for these ads one will typically have uncertainty in the true eCPM of the ad due to the fact that one will not know the click-through rate of a brand new ad with certainty. In this case, it can be desirable to show an ad where one has a high amount of uncertainty about the true eCPM of the ad so one can learn more about the ad's true eCPM by observing whether the ad received a click and using this information to refine one's estimate about the true click-through rate and eCPM of the ad. Thus even if one believes that a high uncertainty ad is not the best ad for this particular query, it may be valuable to show this ad so one can learn more about the eCPM of the ad and make better decisions about whether to show this ad in the future.

While there is an extensive literature that analyzes strategic experimentation in these types of multi-armed bandit problems, the online advertising setting differs substantially from these existing models. In online auctions there is a tremendous amount of random variation in the quality of competition that an ad with unknown eCPM faces in the auction due to the fact that the ad is constantly competing in a wide variety of auctions that may differ in a myriad of different ways. In these settings, there will always be a certain amount of free exploration that takes places due to the fact that there will be some auctions in which there simply are no ads with eCPMs that are known to be high, and one can use these opportunities to explore ads with uncertain eCPMs. Almost all existing models of multi-armed bandits that can be applied to online auctions fail to take this possibility into account.

This paper presents a model of repeated auctions in which an ad with an uncertain click-through rate faces a random distribution of competing bids in each auction and there is discounting of payoffs in the sense that an auctioneer values a dollar received in the distant future less highly than a dollar received today. We formulate this problem as a dynamic programming problem and show that the optimal solution to this problem takes a remarkably simple form. In each period, the auctioneer should rank the advertisers on the basis of the sum of an advertiser's expected eCPM plus a term that represents the value of learning about the eCPM of a particular ad. One then runs the auction by ranking the ads by the ads by these social values rather than their expected eCPMs.

While there have been previous papers on multi-armed bandits that have proposed ranking arms by a term equal to the expected value of showing an ad plus an additional term representing the value of learning about the true value of that arm, the value of learing in the problem that we con-

WWW '14, April 7-11, 2014, Seoul, Korea

sider is dramatically different than the value of learning in standard multi-armed bandit problems. In standard multi-armed problems (*e.g.* Auer *et al.* 2002) where there is no discounting of payoffs and no random variation in the competition that an arm faces in a given period, typical solutions involve ranking the ads according to a sum of the expected value of the arm plus a term proportional to the standard deviation in the arm's value. By contrast, we find that the value of learning in our setting is proportional to the variance in an ad's expected eCPM divided by the number of impressions that an ad has received. Thus the incremental increase in the probability that a particular ad is shown varies with $\frac{1}{k^2}$, where k denotes the number of impressions this ad has received so far. This is an order of magnitude smaller than the corresponding incremental increase in standard machine learning algorithms. In fact, we show that if we attempted to rank the ads on the basis of the sum of an advertiser's expected eCPM plus a term equal to a constant times the standard deviation in the advertiser's eCPM, the optimal constant in such a ranking scheme would be zero.

A consequence of these small incremental changes in the probability that an ad is shown is that the total value from adding active exploration to a machine learning system in the online auction setting is exceedingly small. Not only does the incremental increase in the probability that a particular ad is shown vary with $\frac{1}{k^2}$, but on top of that, the expected payoff increase that one obtains conditional on showing a different ad than would be shown without active learning also varies with $\frac{1}{k^2}$. This implies that the total value of adding active exploration to a machine learning system in the setting we consider will vary with $\frac{1}{k^4}$ for large numbers of impressions k, an exceedingly small amount.

We further obtain finite sample results illustrating that for realistic amounts of uncertainty in the eCPMs of ads with unknown eCPMs, the maximum total efficiency gain that could ever be achieved by adding active learning to a machine learning system in this auction environment is exceedingly small, typically only a few hundredths of a percentage point. Finally, we empirically verify these findings through simulations and illustrate that adding active learning to a machine learning system in the auction environment we consider only changes overall efficiency by a few hundredths of a percentage point.

Perhaps the most closely related paper to our work is [30]. This paper is the only other paper we are aware of that considers questions related to the value of learning about the eCPM bids of ads with uncertain eCPMs in a setting where there is discounting in payoffs as well as random variation in the quality of the competition that an ad faces from competing ads in the auction. In [30], the authors show that the value of showing an ad with an uncertain eCPM will generally exceed the immediate value of showing that ad because one will learn information about the eCPM of the ad that will enable one to make better ranking decisions in the future. However, in [30], the authors do not attempt to characterize the optimal solution in this setting, as we do in the present paper.

There is also an extensive literature in statistics and machine learning that addresses questions related to multi-armed bandits ([5], [6], [22], [28]) as well as some papers that focus specifically on the auction context ([1], [7], [18], [40]). However, none of these papers considers appropriate methods for exploring ads in a context when there is discounting of payoffs, and none of these papers considers appropriate methods for exploring ads when there is random variation in the quality of the competition that an ad faces from competing ads in the auction. The optimal methods for exploring ads in such a scenario turn out to be completely different from any of the methods considered in any of these previous papers, and as such, our work is completely different from existing machine learning literature.

Finally, there is an extensive literature within economics related to questions on strategic experimentation. Within economics, this literature has considered a variety of questions including consumers trying to learn about the quality of various products ([11], [12], [13]), firms trying to learn about the demand curve ([3], [19], [24], [31], [35]), learning to play repeated games ([4], [21]), learning about untried policies in a political economy setting ([17], [37]), learning from the actions of others ([8], [20], [38]), as well as general theoretical results on experimentation ([2], [9], [14], [15], [16], [25], [26], [32], [34], [36], [39]). However, the economics literature has not considered strategic experimentation in the online auction setting, as we do in the present paper.

2. THE MODEL

There is a new ad with an uncertain eCPM that will bid into an auction with competing advertisers. Throughout we let x denote the actual, unknown value for showing the new ad and z denote the bid that this ad places in the auction (on an eCPM scale). We also let k denote the number of impressions the ad has received so far. Finally we assume that the ad has some underlying type θ^* in the set Θ, where Θ denotes the set of all possible types of the ad. One can think of θ^* as representing all possible qualities of the ad that are relevant towards determining the eCPM of the ad such as the clickability or the quality score of the ad.

For any fixed type of the ad, θ^*, there will be an associated eCPM of the ad. We allow for the possibility that, even if the underlying type of the ad is known and fixed, the eCPM of the ad may evolve over time as the ad is shown. This is relevant, for instance, with wear-out of ads. For certain ads, even if the underlying type or quality of the ad is known with certainty, it is possible that this ad will lose its effectiveness over time if the ad is shown over and over again because users become used to the ad and are less inclined to click on the ad than before. Thus for any fixed type of the ad, θ^*, there is some associated eCPM of the ad $x(\theta^*, k)$ that depends both on the underlying type of the ad as well as the number of impressions the ad has received so far.

While our model allows for the possibility of wear-out of an ad, empirical studies of wear-out in online advertising suggest that this wear-out is exceedingly small for ads that have already been shown a large number of times [29]. In our paper, we model this by assuming that $x(\theta^*, k) - x(\theta^*, k+1) = o(1/k)$ for all possible underlying types of the ad, θ^*. Our formulation also implicitly assumes that the number of impressions an ad has received can influence the ad's eCPM, but not when these impressions occurred. While this assumption may not be perfect in every situation, it seems reasonable in situations where overexposure is the reason that the number of impressions an ad has received influneces the ad's eCPM, and it is a standard assumption in the multi-armed bandits literature.

At any given point in time, the auctioneer does not necessarily observe the exact type of the ad. Instead the auction-

eer only knows that this type is drawn from some distribution. We let $\tilde{\theta}$ denote a generic distribution corresponding to the auctioneer's estimate of the distribution of types that the ad may assume. This distribution will evolve over time as an ad has received more impressions and we have a better sense of the underlying eCPM of the ad. We sometimes write $\tilde{\theta}_k$ to denote the auctioneer's estimate of the distribution of types for the ad after the ad has been shown k times.

Throughout we also let $\overline{x}(\tilde{\theta}, k)$ denote an unbiased estimate of the true value of x given the estimated distribution of types for the ad, $\tilde{\theta}$, and the number of impressions that the ad has received, k. This value of $\overline{x}(\tilde{\theta}, k)$ is just equal to the integral of the possible values of $x(\theta^*, k)$ weighted by the relative likelihoods that the type of the ad is θ^* in the distribution $\tilde{\theta}$.

We also let σ_k^2 denote the variance in our estimate of the eCPM for the new ad when the ad has been shown k times. In the limit when k is large, σ_k^2 will be well approximated by $\frac{s^2(\overline{x}(\tilde{\theta},k))}{k}$ for some constant $s^2(\overline{x}(\tilde{\theta}, k))$ that depends only on $\overline{x}(\tilde{\theta}, k)$. In addition, we let $\delta \in (0, 1)$ denote the per-period discount rate so that the mechanism designer only values advertising opportunities that take place at time T by a factor of δ^T as much as opportunities that take place at the present time period.

In the model we consider we restrict attention to settings in which there is a single advertising opportunity that is being sold at an auction, and the auction is being conducted using a second price format. We suppose throughout that the distribution of the values of the competing advertisers is such that the highest eCPM for a competing ad is a random draw from some cumulative distribution function $F(\cdot)$ with corresponding density $f(\cdot)$.

3. DYNAMIC PROGRAMMING PROBLEM

In this section we formulate the value of a particular ad as a dynamic programming problem and use this formulation to derive the optimal bidding strategy for a particular ad. First we derive the total social value that arises in a particular period when a new ad makes a particular bid.

Note that if the new ad places a bid of z in the auction and the actual value of showing this particular ad is x, then the total social welfare that arises as a result of running the auction once is

$$\begin{aligned} u &= \int_z^\infty y f(y)\, dy + \int_0^z x f(y)\, dy \\ &= -y(1-F(y))|_z^\infty + \int_z^\infty (1-F(y))\, dy + xF(z) \\ &= z(1-F(z)) + \int_z^\infty (1-F(y))\, dy + xF(z) \end{aligned}$$

In general placing a bid of z rather than x in a one-shot auction will result in some inefficencies in the one-shot auction since it would be optimal for social welfare if the new ad placed a bid exactly equal to x in a one-shot auction. If $u(z, x)$ denotes the total social welfare that arises when a new ad with value x places an eCPM bid of z, then the social loss that arises in a one-shot auction as a result of placing a bid of z instead of x is

$$\begin{aligned} L &= u(x,x) - u(z,x) \\ &= x(1-F(x)) + \int_x^\infty (1-F(y))\, dy + xF(x) \\ &\quad -z(1-F(z)) - \int_z^\infty (1-F(y))\, dy - xF(z) \\ &= (x-z)(1-F(z)) + \int_x^z (1-F(y))\, dy \\ &= \int_x^z F(z) - F(y)\, dy \end{aligned}$$

Now let $V_k(\overline{x}(\tilde{\theta}_k, k))$ denote the value of the dynamic program from displaying an ad that has an expected value of $\overline{x}(\tilde{\theta}_k, k)$ if the ad has been shown k times. Note that we can express x as $x = \overline{x}(\tilde{\theta}_k, k) + \sigma_k \epsilon$, where σ_k denotes the standard deviation in our estimate of the eCPM for the new ad when the ad has been shown k times, and ϵ is a random variable with mean zero and variance one. We use this notation in proving the following result:

THEOREM 1. *The value of the dynamic programming problem can be expressed as* $V_k(\overline{x}(\tilde{\theta}_k, k)) =$

$$\frac{1}{1-\delta}\left(\max_z E_\epsilon\left[-\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^z F(z) - F(y)\, dy + \right.\right.$$
$$\left.\left.\delta F(z)(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] - V_k(\overline{x}(\tilde{\theta}_k, k)))\right]\right)$$

PROOF. Suppose an ad has been shown k times and has an estimated distribution of types $\tilde{\theta}_k$ and expected value of x equal to $\overline{x}(\tilde{\theta}_k, k)$. The value of the dynamic programming problem that arises from placing the optimal bid z in the current period, $V_k(\overline{x}(\tilde{\theta}_k, k))$, is equal to the immediate reward from bidding z (or the negative of the loss function) that arises in the current period plus δ times the expected value of the dynamic programming problem that arises in the next period.

Now if the new advertiser places a bid of z, then the probability the advertiser wins the auction is $F(z)$, in which case the expected value of the dynamic programming problem that arises next period is $E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))]$, where the expectation is taken over the randomness in the changes in the estimates of the distribution of types $\tilde{\theta}$ that arise as a result of showing this ad. The probability the advertiser does not win the auction is $1 - F(z)$, in which case the value of the dynamic programming problem that arises next period remains at $V_k(\overline{x}(\theta_k, k))$. Thus the expected value of the dynamic programming problem that arises in the next period is $F(z)E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] + (1-F(z))V_k(\overline{x}(\tilde{\theta}_k, k)))$.

At the same time we have already seen that the social value from bidding z that arises in the current period equals $-\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^z F(z) - F(y)\, dy$. By combining this with the insights in the previous paragraphs, it follows that $V_k(\overline{x}(\tilde{\theta}_k, k)) =$

$$\max_z E_\epsilon\left[-\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^z F(z) - F(y)\, dy + \right.$$
$$\left.\delta(F(z)E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] + (1-F(z))V_k(\overline{x}(\tilde{\theta}_k, k)))\right]$$

By subtracting $\delta V_k(\overline{x}(\tilde{\theta}_k, k))$ from both sides and dividing both sides by $1 - \delta$, it then follows that $V_k(\overline{x}(\tilde{\theta}_k, k)) =$

$$\frac{1}{1-\delta}\left(\max_z E_\epsilon\left[-\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z)-F(y)\ dy+ \delta F(z)(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))\right]\right)$$

□

By using the expression for the value of the dynamic programming problem in the previous theorem, we can derive the bid that an advertiser should place to maximize social welfare in this setting. This is done in the theorem below:

THEOREM 2. *The optimal bidding strategy in the dynamic programming problem when an ad has been shown k times, has an estimated distribution of types $\tilde{\theta}_k$, and has an expected value of x equal to $\overline{x}(\tilde{\theta}_k,k)$ entails setting $z=\overline{x}(\tilde{\theta}_k,k)+\delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))$.*

PROOF. By differentiating the expression in Theorem 1 with respect to z, we see that the first order condition for z to be an optimal bid is

$$\begin{aligned}0 &= E_\epsilon\left[-\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} f(z)\ dy+ \delta f(z)(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))\right]\\ &= E_\epsilon\left[-f(z)(z-\overline{x}(\tilde{\theta}_k,k)-\sigma_k\epsilon)+ \delta f(z)(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))\right]\\ &= f(z)(\overline{x}(\tilde{\theta}_k,k)-z+ \delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k))))\end{aligned}$$

From this it follows that the optimal bidding strategy in the dynamic programming problem entails setting $z=\overline{x}(\tilde{\theta}_k,k)+\delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))$. □

Thus the optimal bidding strategy in this dynamic programming problem can be written in the form whereby the bidder with uncertain eCPM makes a bid equal to the bidder's expected eCPM plus a term that represents the value of learning about the true eCPM of that bidder, $\delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))$. In order to calculate this value of learning, we need to get a sense of the size of the $V_k(\overline{x}(\tilde{\theta}_k,k))$ terms.

4. VALUE OF DYNAMIC PROGRAM FOR LARGE NUMBERS OF IMPRESSIONS

In the previous section we have given exact expressions for the value of the dynamic programming problem and the optimal bidding strategy that should be followed under this dynamic programming problem. In this section, we seek to derive accurate estimates of the value of this dynamic programming problem in the limit when an ad has already been shown a large number of times.

When an ad has already been shown a large number of times, the value of σ_k that is estimated for the ad is likely to be very small. For small values of σ_k, we can use a Taylor expansion to approximate the value of the above dynamic programming problem. In particular, we obtain the following result:

THEOREM 3. *$E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z)-F(y)\ dy]=\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z)-F(y)\ dy+\frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k,k))+o(\sigma_k^2)$ for large k.*

PROOF. If $J(\sigma_k)=E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z)-F(y)\ dy]$, then $J(0)=\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z)-F(y)\ dy$, $J'(\sigma_k)=-E_\epsilon[\epsilon(F(z)-F(\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon))]$, $J'(0)=0$, and $J''(0)=E_\epsilon\epsilon^2 f(\overline{x}(\tilde{\theta}_k,k))=f(\overline{x}(\tilde{\theta}_k,k))$. From this it follows that the second-order Taylor approximation to $E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z)-F(y)\ dy]$ is $\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z)-F(y)\ dy+\frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k,k))+o(\sigma_k^2)$. □

Using the results from the previous theorem, one can immediately illustrate that V_k must be on the order of $\frac{1}{k}$ for large values of k.

THEOREM 4. *$V_k(\overline{x}(\tilde{\theta}_k,k))=\Theta(\frac{1}{k})$ for large k.*

PROOF. First note that it must be the case that $V_k(\overline{x}(\tilde{\theta}_k,k))=\Omega(\frac{1}{k})$ for large k. We know that $\sigma_k^2=\Theta(\frac{1}{k})$ for large k, and we also know from the expression in the previous theorem that the immediate reward in any given period is at least on the same order as $\frac{1}{k}$. Thus we know that $V_k(\overline{x}(\tilde{\theta}_k,k))=\Omega(\frac{1}{k})$ for large k. But we also know that $V_k(\overline{x}(\tilde{\theta}_k,k))=O(\frac{1}{k})$ for large k. To see this, note that the auctioneer can ensure that his loss in any given period is $O(\frac{1}{k})$ by bidding $z=\overline{x}(\tilde{\theta}_k,k)$. And if the auctioneer's loss in any given period is $O(\frac{1}{k})$, then the auctioneer's total loss from the game will also be no greater than $O(\frac{1}{k})$ because the present value of the sum of losses that are $\Theta(\frac{1}{k})$, $\sum_{j=k}^{\infty}\delta^{j-k}\frac{v}{j}$, is also $\Theta(\frac{1}{k})$ since $1<\sum_{j=k}^{\infty}\delta^{j-k}\frac{k}{j}<\sum_{j=k}^{\infty}\delta^{j-k}=\frac{1}{1-\delta}$ implies $\frac{1}{k}<\sum_{j=k}^{\infty}\delta^{j-k}\frac{1}{j}<\frac{1}{(1-\delta)k}$. Thus $V_k(\overline{x}(\tilde{\theta}_k,k))=\Theta(\frac{1}{k})$ for large k. □

To understand the intuition behind this result, note that the average error in the estimate of the eCPM of the ad is proportional to the standard error of this estimate, σ_k, which varies with $\frac{1}{\sqrt{k}}$, so the probability that the auctioneer will display the wrong ad as a result of misestimating the eCPM of the ad varies with $\frac{1}{\sqrt{k}}$. At the same time, conditional on displaying the wrong ad as a result of misestimating the eCPM of the ad, the average efficiency loss that one suffers varies with $\frac{1}{\sqrt{k}}$. Thus the expected efficiency loss that the auctioneer incurs varies with $\frac{1}{k}$, which in turn implies the result in Theorem 4.

Theorem 4 suggests that we may be able to express $V_k(\overline{x}(\tilde{\theta}_k,k))$ by $V_k(\overline{x}(\tilde{\theta}_k,k))=-\frac{v(\overline{x}(\tilde{\theta}_k,k))}{k}+o(\frac{1}{k})$ for large k, where v is a function that depends only on $\overline{x}(\tilde{\theta}_k,k)$. To prove that $V_k(\overline{x}(\tilde{\theta}_k,k))$ can be expressed this way, it is necessary to show that $kV_k(\overline{x}(\tilde{\theta}_k,k))$ indeed converges to a function of $\overline{x}(\tilde{\theta}_k,k)$ in the limit as $k\to\infty$. This is done in the following theorem:

THEOREM 5. *$kV_k(\overline{x}(\tilde{\theta}_k,k))$ converges to a function of $\overline{x}(\tilde{\theta}_k,k)$ in the limit as $k\to\infty$. Furthermore, it must be the case that $\lim_{k\to\infty} kV_k(\overline{x}(\tilde{\theta}_k,k))=-\frac{1}{2(1-\delta)}s^2(\overline{x}(\tilde{\theta}_k,k))f(\overline{x}(\tilde{\theta}_k,k))$.*

PROOF. Since $V_k(\overline{x}(\tilde{\theta}_k,k))=\Theta(\frac{1}{k})$ for large k, it must be the case that $E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k))=$

$O(\frac{1}{k})$ for large k. Thus since the optimal bidding strategy entails setting $z = \overline{x}(\tilde{\theta}_k, k) + \delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] - V_k(\overline{x}(\tilde{\theta}_k, k)))$, it must be the case that $z = \overline{x}(\tilde{\theta}_k, k) + O(\frac{1}{k})$ for large k. From this it follows that $\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z) - F(y)\ dy = O(\frac{1}{k^2})$ under the optimal bidding strategy z for large k.

Now we have seen in Theorem 3 that $E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z) - F(y)\ dy + \frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k, k)) + o(\sigma_k^2)$ for large k. Since we know that $\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z) - F(y)\ dy = O(\frac{1}{k^2})$ under the optimal bidding strategy z and we have assumed that $\sigma_k^2 = \frac{s^2(\overline{x}(\tilde{\theta},k))}{k} + o(\frac{1}{k})$ for large k, it then follows that $E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = \frac{1}{2k} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + o(\frac{1}{k})$ for large k.

But $-E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy]$ represents the per-period utility that one obtains at each point in the game. Since $V_k(\overline{x}(\tilde{\theta}_k, k))$ can alternatively be expressed as the discounted sum of the per-period utility that one can obtain at each point in the game, it then follows that $|kV_k(\overline{x}(\tilde{\theta}_k, k))| \leq \sum_{j=k}^{\infty} \delta^{j-k}[\frac{1}{2} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k))] + o(1)$, meaning $|kV_k(\overline{x}(\tilde{\theta}_k, k))| \leq \frac{1}{2(1-\delta)} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + o(1)$ and $|kV_k(\overline{x}(\tilde{\theta}_k, k))| \geq \sum_{j=k}^{\infty} \delta^{j-k}[\frac{k}{2j} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k))] + o(1) = \frac{1}{2(1-\delta)} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + o(1)$ in the limit as $k \to \infty$. From this it follows that $|kV_k(\overline{x}(\tilde{\theta}_k, k))| = \frac{1}{2(1-\delta)} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{t}_k, k)) + o(1)$ and $\lim_{k\to\infty} kV_k(\overline{x}(\tilde{\theta}_k, k)) = -\frac{1}{2(1-\delta)} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k))$. □

From Theorem 5, it follows that we can express $V_k(\overline{x}(\tilde{\theta}_k, k))$ by $V_k(\overline{x}(\tilde{\theta}_k, k)) = -\frac{v(\overline{x}(\tilde{\theta}_k,k))}{k} + o(\frac{1}{k})$ for large k, where v is a function that satisfies $v(\overline{x}(\tilde{\theta}_k, k)) = \frac{1}{2(1-\delta)} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k))$. In order to complete our approximation of the solution the dynamic programming problem for large k, it is also necessary to bound the expression $E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] - V_k(\overline{x}(\tilde{\theta}_k, k))$ that appears in the dynamic programming problem. This is done in the following theorem:

THEOREM 6. *$E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1}, k+1))] - V_k(\overline{x}(\tilde{\theta}_k, k)) = \frac{v(\overline{x}(\tilde{\theta}_k,k))}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k.*

PROOF. Note that if an ad is displayed, then one of two possible things will happen to the ad—either the ad will receive a click or the ad will not receive a click. Let p denote the probability that the ad will receive a click, let $\tilde{\theta}_c$ denote the estimated distribution of types for the ad if the ad receives a click, and let $\tilde{\theta}_n$ denote the estimated distribution of types for the ad if the ad does not receive a click. Note that if $\tilde{\theta}_k$ denotes the estimated distribution of types for the ad before the ad was displayed, then it must be the case that $p\tilde{\theta}_c + (1-p)\tilde{\theta}_n = \tilde{\theta}_k$. And if $\overline{x}_c \equiv \overline{x}(\tilde{\theta}_c, k+1)$, $\overline{x}_n \equiv \overline{x}(\tilde{\theta}_n, k+1)$, and $\overline{x} \equiv \overline{x}(\tilde{\theta}_k, k+1)$, then it also must be the case that $p\overline{x}_c + (1-p)\overline{x}_n = \overline{x}$.

Now note that the second-order Taylor approximations for $V_{k+1}(\overline{x}_c)$ and $V_{k+1}(\overline{x}_n)$ are

$$V_{k+1}(\overline{x}_c) \approx V_{k+1}(\overline{x}) + V'_{k+1}(\overline{x})(\overline{x}_c - \overline{x}) + \frac{1}{2} V''_{k+1}(\overline{x})(\overline{x}_c - \overline{x})^2$$

and

$$V_{k+1}(\overline{x}_n) \approx V_{k+1}(\overline{x}) + V'_{k+1}(\overline{x})(\overline{x}_n - \overline{x}) + \frac{1}{2} V''_{k+1}(\overline{x})(\overline{x}_n - \overline{x})^2$$

Thus if $\overline{x}'$ denotes the actual realization of the estimated eCPM after the ad has been shown $k+1$ times ($\overline{x}'$ will equal $\overline{x}_c$ with probability p and $\overline{x}_n$ with probability $1-p$), then by utilizing the fact that $p\overline{x}_c + (1-p)\overline{x}_n = \overline{x}$ and by taking a weighted average of the two previous equations, we find that

$$\begin{aligned} E[V_{k+1}(\overline{x}')] &= pV_{k+1}(\overline{x}_c) + (1-p)V_{k+1}(\overline{x}_n) \\ &\approx V_{k+1}(\overline{x}) + \frac{1}{2} V''_{k+1}(\overline{x}) E[(\overline{x}' - \overline{x})^2] \end{aligned}$$

From this it follows that $E[V_{k+1}(\overline{x}') - V_k(\overline{x}(\tilde{\theta}_k, k))]$ is

$$\begin{aligned} &\approx V_{k+1}(\overline{x}) - V_k(\overline{x}(\tilde{\theta}_k, k)) + \frac{1}{2} V''_{k+1}(\overline{x}) E[(\overline{x}' - \overline{x})^2] \\ &\approx \frac{v(\overline{x}(\tilde{\theta}_k, k))}{k} - \frac{v(\overline{x}(\tilde{\theta}_k, k+1))}{k+1} - \frac{v''(\overline{x}(\tilde{\theta}_k, k+1))}{2(k+1)} E[(\overline{x}' - \overline{x})^2] \\ &\approx \frac{(k+1)v(\overline{x}(\tilde{\theta}_k, k)) - kv(\overline{x}(\tilde{\theta}_k, k+1))}{k(k+1)} - \frac{v''(\overline{x}(\tilde{\theta}_k, k+1))}{2(k+1)} E[(\overline{x}' - \overline{x})^2] \\ &= \frac{v(\overline{x}(\tilde{\theta}_k, k))}{k(k+1)} - \frac{v(\overline{x}(\tilde{\theta}_k, k)) - v(\overline{x}(\tilde{\theta}_k, k+1))}{k+1} \\ &\quad - \frac{v''(\overline{x}(\tilde{\theta}_k, k+1))}{2(k+1)} E[(\overline{x}' - \overline{x})^2] \end{aligned}$$

If c denotes the number of clicks that an ad has received so far, then the predicted click-through rate for an ad that has received a large number of impressions, k, will be approximately $\frac{c}{k}$. Thus if b denotes the bid per click that the ad places, then the eCPM for an ad that has received c clicks and has been shown k times will be $\overline{x} \approx \frac{bc}{k}$. From this it follows that $\overline{x}_c \approx \frac{b(c+1)}{k+1}$, $\overline{x}_n \approx \frac{bc}{k+1}$, $\overline{x}_c - \overline{x} \approx \frac{b(k-c)}{k(k+1)}$, and $\overline{x}_n - \overline{x} \approx -\frac{bc}{k(k+1)}$. Thus $\overline{x}' - \overline{x} = O(\frac{1}{k})$ for all possible realizations of $\overline{x}'$, and $(\overline{x}' - \overline{x})^2 = O(\frac{1}{k^2})$ for all possible realizations of $\overline{x}'$ as well. We thus know that

$$\frac{v''(\overline{x}(\tilde{\theta}_k, k+1))}{2(k+1)} E[(\overline{x}' - \overline{x})^2] = O\left(\frac{1}{k^3}\right) \tag{1}$$

Now it must be the case that $\overline{x}(\tilde{\theta}_k, k) - \overline{x}(\tilde{\theta}_k, k+1) = o(\frac{1}{k})$ because $x(\theta^*, k) - x(\theta^*, k+1) = o(1/k)$ for all $\theta^* \in \Theta$ by assumption. Thus $\overline{x}(\tilde{\theta}_k, k) - \overline{x}(\tilde{\theta}_k, k+1) = o(\frac{1}{k})$, and it also follows that

$$v(\overline{x}(\tilde{\theta}_k, k)) - v(\overline{x}(\tilde{\theta}_k, k+1)) = o\left(\frac{1}{k}\right) \tag{2}$$

as well.

By using the results in equations (1) and (2), we see that it must be the case that $\frac{v(\overline{x}(\tilde{\theta}_k,k)) - v(\overline{x}(\tilde{\theta}_k,k+1))}{k+1} = o(\frac{1}{k^2})$ and $\frac{v''(\overline{x}(\tilde{\theta}_k,k))}{2(k+1)} E[(\overline{x}' - \overline{x})^2] = o(\frac{1}{k^2})$ as well. Substituting these results in to our earlier approximation for the value of $E[V_{k+1}(\overline{x}') - V_k(\overline{x}(\tilde{\theta}_k, k))]$ then gives

$$E[V_{k+1}(\overline{x}') - V_k(\overline{x}(\tilde{\theta}, k))] = \frac{v(\overline{x}(\tilde{\theta}, k))}{k(k+1)} + o\left(\frac{1}{k^2}\right)$$

□

The intuition behind this result is that since the efficiency loss that the auctioneer incurs from uncertainty in the eCPM

of an ad varies with $\frac{1}{k}$, the value of learning will be proportional to the reduction in the future efficiency loss that the auctioneer suffers as a result of learning more about the eCPM of the ad, meaning the value of learning will vary with $\frac{1}{k} - \frac{1}{k+1}$, which varies with $\frac{1}{k^2}$. The fact that $E[V_{k+1}(\overline{x}') - V_k(\overline{x}(\tilde{\theta}_k, k))] = \frac{v(\overline{x}(\tilde{\theta}_k,k))}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ varies with $\frac{1}{k^2}$ indicates that the incremental increase in an advertiser's bid also varies with $\frac{1}{k^2}$ in the limit when k is large. This in turn also implies that the incremental increase in an advertiser's probability of winning the auction will also vary with $\frac{1}{k^2}$ in the limit when k is large.

The result in Theorem 6 suggests that the optimal method for adding active exploration into a machine learning system in online auctions will only rarely have an effect on which ad wins the auction, as the probability that this active exploration changes which ad is shown varies with $\frac{1}{k^2}$ for large k. This result about the value of learning varying with $\frac{1}{k^2}$ for large k stands in marked contrast to algorithms that have been proposed for active exploration in standard multi-armed bandit problems with no discounting of payoffs and no random variation in the competition that an arms faces in a given period (e.g. [5]). In these types of algorithms, the value of learning tends to vary with $\frac{1}{\sqrt{k}}$, which means the value of learning is an order of magnitude smaller in our setting than in standard multi-armed bandit problems. Thus the value of learning is dramatically different in an online auction setting than in a standard multi-armed bandit problem.

5. PERFORMANCE GUARANTEES

The results in the previous sections suggest a possible algorithm that will well approximate the optimal bidding strategies for an auctioneer who seeks to show the advertisement that will lead to the greatest social welfare, where this welfare includes the value of learning about the eCPM's of the advertisers with unknown eCPM's. This algorithm would proceed by computing the expected eCPM for an advertiser with unknown eCPM, $\overline{x}$, the density for the distribution of competing eCPM bids at this value of $\overline{x}$, $f(\overline{x})$, the variance $s^2(\overline{x})$ in the eCPM for an ad with estimated eCPM $\overline{x}$ that has only received one impression, and the number of impressions k that the ad has received. One then decides which ads to show by computing a score equal to $\overline{x} + \frac{\delta}{2(1-\delta)k(k+1)} s^2(\overline{x}) f(\overline{x})$ for each of the advertisers, where δ is the auctioneer's discount factor, and showing the ad from the advertiser with the highest such score. In this section we address questions related to the payoffs that the auctioneer can obtain by using this algorithm and related algorithms.

First we address questions related to how the algorithms we have considered in this paper will compare to other plausible algorithms that have been considered in the machine learning literature. One other algorithm that is standard for multi-armed bandit problems in the machine learning literature is an algorithm which involves ranking the arms by a term equal to the expected value of the arm plus a term that is proportional to the standard deviation in the arm [5]. More generally, one can rank advertisers by a term equal to the eCPM of the advertiser plus a term that is proportional to $\frac{1}{k^\alpha}$ for any $\alpha \leq \frac{1}{2}$, where k denotes the number of impressions that the ad has received so far. However, these algorithms are not well-suited towards the auction environment, as the following theorem illustrates:

THEOREM 7. *Suppose the auctioneer uses a bid for the advertiser with unknown eCPM that is of the form* $z = \overline{x}(\tilde{\theta}_k, k) + \frac{c(\overline{x}(\tilde{\theta}_k,k))}{k^\alpha}$, *where* $c(\overline{x}(\tilde{\theta}_k, k))$ *is a bounded non-negative constant that depends only on the term* $\overline{x}(\tilde{\theta}_k, k)$ *(and the distribution of competing bids), and* $\alpha \leq \frac{1}{2}$. *Then the optimal constant* $c(\overline{x}(\tilde{\theta}_k, k))$ *for any such algorithm is* $c(\overline{x}(\tilde{\theta}_k, k)) = 0$ *for sufficiently large* k.

PROOF. Recall from the proof of Theorem 3 that the auctioneer's per-period payoff if the auctioneer uses a bid for the advertiser with unknown eCPM that is equal to z is $-E_\epsilon[\int_{\overline{x}(\tilde{\theta}_k,k)+\sigma_k\epsilon}^{z} F(z) - F(y)\ dy] = -\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z) - F(y)\ dy - \frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k, k)) + o(\sigma_k^2)$ for large k. Now if $z = \overline{x}(\tilde{\theta}_k, k) + \frac{c(\overline{x}(\tilde{\theta}_k,k))}{k^\alpha}$ for some constant $c(\overline{x}(\tilde{\theta}_k, k))$, then $\int_{\overline{x}(\tilde{\theta}_k,k)}^{z} F(z) - F(y)\ dy = \int_{\overline{x}(\tilde{\theta}_k,k)}^{\overline{x}(\tilde{\theta}_k,k)+\frac{c(\overline{x}(\tilde{\theta}_k,k))}{k^\alpha}} f(\overline{x}(\tilde{\theta}_k, k))(\overline{x}(\tilde{\theta}_k, k) + \frac{c(\overline{x}(\tilde{\theta}_k,k))}{k^\alpha} - y)\ dy + o(\frac{1}{k^{2\alpha}}) = f(\overline{x}(\tilde{\theta}_k, k))\frac{c^2(\overline{x}(\tilde{\theta}_k,k))^2}{2k^{2\alpha}}$. Thus the auctioneer's per-period payoff is if the auctioneer uses a bid for the advertiser with unknown eCPM of the form $z = \overline{x}(\tilde{\theta}_k, k) + \frac{c(\overline{x}(\tilde{\theta}_k,k))}{k^\alpha}$ is $-\frac{c^2(\overline{x}(\tilde{\theta}_k,k))}{2k^{2\alpha}} - \frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k, k)) + o(\sigma_k^2)$.

Now if $c(\overline{x}(\tilde{\theta}_k, k)) = 0$, then the auctioneer's per-period payoff is $-\frac{1}{2}\sigma_k^2 f(\overline{x}(\tilde{\theta}_k, k)) + o(\sigma_k^2) = -\frac{1}{2k} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + o(\frac{1}{k})$. We then know from the reasoning in the proof of Theorem 5 that if this is the auctioneer's per-period payoff, then the auctioneer's total payoff from the game is $-\frac{1}{2(1-\delta)k} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k))) + o(\frac{1}{k})$ regardless of the learning rate. Similarly, if $c(\overline{x}(\tilde{\theta}_k, k)) \neq 0$ and $\alpha = \frac{1}{2}$, then the auctioneer's per-period payoff is $-\frac{1}{2k}(s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + c^2(\overline{x}(\tilde{\theta}_k, k))) + o(\frac{1}{k})$, and we know from identical reasoning that the auctioneer's total payoff from the game is $-\frac{1}{2(1-\delta)k}(s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + c^2(\overline{x}(\tilde{\theta}_k, k))) + o(\frac{1}{k})$, which is strictly less than the auctioneer's total payoff from the game when $c(\overline{x}(\tilde{\theta}_k, k)) = 0$ for sufficiently large k.

Finally, if $c(\overline{x}(\tilde{\theta}_k, k)) \neq 0$ and $\alpha < \frac{1}{2}$, then the auctioneer's per-period payoff is $-\frac{c^2(\overline{x}(\tilde{\theta}_k,k))}{2k^{2\alpha}} + o(\frac{1}{k^{2\alpha}})$. Since the auctioneer's payoff from the game is equal to the discounted sum of the auctioneer's per-period payoffs, it then follows that if V_k denotes the auctioneer's total payoff from the game from using this strategy, then $k^{2\alpha} V_k(\overline{x}(\tilde{\theta}_k, k)) \leq \sum_{j=k}^{\infty} \delta^{j-k}[-\frac{1}{2}(\frac{k}{j})^{2\alpha} c^2(\overline{x}(\tilde{\theta}_k, k))] + o(1) = -\frac{1}{2(1-\delta)} c^2(\overline{x}(\tilde{\theta}_k, k)) + o(1)$ in the limit as $k \to \infty$. Thus if $c(\overline{x}(\tilde{\theta}_k, k)) \neq 0$ and $\alpha < \frac{1}{2}$, then the auctioneer's payoff from the game is no greater than $-\frac{1}{2(1-\delta)k^{2\alpha}} c^2(\overline{x}(\tilde{\theta}_k, k)) + o(\frac{1}{k^{2\alpha}})$, which is less than $-\frac{1}{2k} s^2(\overline{x}(\tilde{\theta}_k, k)) f(\overline{x}(\tilde{\theta}_k, k)) + o(\frac{1}{k})$, the auctioneer's payoff from using the constant $c(\overline{x}(\tilde{\theta}_k, k)) = 0$ for sufficiently large k. From this and the result in the previous paragraph it follows that if the auctioneer is using the strategy given in the statement of this theorem, the auctioneer's total payoff for the game will be maximized when $c(\overline{x}(\tilde{\theta}_k, k)) = 0$ for sufficiently large k. □

This result immediately implies that standard existing machine learning algorithms for exploration which involve adding a term proportional to the standard deviation to the eCPM of the ad, such as the UCB algorithm, are actually dominated by the simple greedy approach of just always making a bid equal to the eCPM of the ad in an auction

environment with discounting of payoffs. These existing algorithms do too much exploration, and as a result of this, lead to lower payoffs than the simple approach of not doing any active exploration at all.

Next we turn to the question of what guarantees can be given about the size of the performance improvement that could be obtained by using the algorithm we have proposed rather than the simple greedy algorithm. Our next result illustrates that one will indeed obtain a performance improvement by using the algorithm that we have proposed, but the size of the performance improvement is likely to be very small.

THEOREM 8. *Suppose the auctioneer uses the algorithm we have outlined. Then the expected payoff that the auctioneer will obtain by using this algorithm will exceed the expected payoff that the auctioneer would obtain by using the purely greedy approach by an amount* $\frac{\delta^2}{8(1-\delta)^3k^4}s^4(\overline{x}(\tilde{\theta}_k,k))f^3(\overline{x}(\tilde{\theta}_k,k))+o(\frac{1}{k^4})$.

PROOF. We know from Theorem 6 that $E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)) = \frac{v(\overline{x}(\tilde{\theta}_k,k))}{k(k+1)} + o\left(\frac{1}{k^2}\right)$ for large k, where $v(\overline{x}(\tilde{\theta}_k,k)) = \frac{1}{2(1-\delta)}s^2(\overline{x}(\tilde{\theta}_k,k))f(\overline{x}(\tilde{\theta}_k,k))$, and we also know from the proof of Theorem 2 that the derivative of the seller's expected payoff from making a bid of z with respect to z is $f(z)(\overline{x}(\tilde{\theta}_k,k)-z+\delta(E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k))))$. Thus if we let $\Delta V \equiv E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)))$, then the difference between the auctioneer's expected payoff from making a bid of $\overline{x}(\tilde{\theta}_k,k))$ and the auctioneer's expected payoff from making a bid of $\overline{x}(\tilde{\theta}_k,k) + \frac{\delta}{2(1-\delta)k(k+1)}s^2(\overline{x}(\tilde{\theta}_k,k))f(\overline{x}(\tilde{\theta}_k,k))$ is

$$\int_{\overline{x}(\tilde{\theta}_k,k))}^{\overline{x}(\tilde{\theta}_k,k)+\delta\Delta V+o(\Delta V)} \frac{f(z)(\overline{x}(\tilde{\theta}_k,k)-z+\delta(\Delta V)+o(\Delta V))}{1-\delta}\,dz$$
$$= \frac{f(\overline{x}(\tilde{\theta}_k,k))\delta^2(\Delta V)^2}{2(1-\delta)} + o((\Delta V)^2)$$

And since $\Delta V = E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k)) = \frac{v(\overline{x}(\tilde{\theta}_k,k))}{k(k+1)}+o\left(\frac{1}{k^2}\right) = \frac{s^2(\overline{x}(\tilde{\theta}_k,k))f(\overline{x}(\tilde{\theta}_k,k))}{2(1-\delta)k(k+1)}+o\left(\frac{1}{k^2}\right)$, it then follows that the difference between the auctioneer's expected payoff from making a bid of $\overline{x}(\tilde{\theta}_k,k))$ and the auctioneer's expected payoff from making a bid of $\overline{x}(\tilde{\theta}_k,k) + \frac{\delta}{2(1-\delta)k(k+1)}s^2(\overline{x}(\tilde{\theta}_k,k))f(\overline{x}(\tilde{\theta}_k,k))$ is $\frac{\delta^2}{8(1-\delta)^3k^4}s^4(\overline{x}(\tilde{\theta}_k,k))f^3(\overline{x}(\tilde{\theta}_k,k)) + o(\frac{1}{k^4})$. The result then follows. □

Theorem 8 indicates that the performance improvement that can be obtained as a result of using the algorithm that we have suggested is only on the order of $\frac{1}{k^4}$, where k denotes the number of impressions that an ad has received. This follows from the fact that the incremental increase in the probability that a particular ad is shown varies with $\frac{1}{k^2}$, and on top of that, the expected payoff increase that one obtains conditional on showing a different ad than would be shown without active learning also varies with $\frac{1}{k^2}$. Since this represents a fourth-order improvement in performance relative to the purely greedy approach, this result indicates that the performance improvement that can be obtained by following our algorithm rather than simply ranking the ads by their eCPM's becomes small very quickly.

It is worth noting, however, that the result in Theorem 8 is not due to our algorithm being a suboptimal implementation of incorporating active exploration into a machine learning system. Our next result illustrates that while the size of the performance improvement that can be obtained by from using our algorithm is small, this algorithm will, in fact, obtain nearly the maximum possible performance improvement over the purely greedy approach of ranking ads by their eCPM's.

THEOREM 9. *Suppose the auctioneer uses the algorithm we have outlined. Then the difference between the auctioneer's payoff under this strategy and the maximum possible payoff the auctioneer could obtain under the theoretically optimal strategy becomes vanishingly small compared to the difference between the auctioneer's payoff under this strategy and the auctioneer's payoff under the greedy strategy for large k.*

PROOF. The theoretically optimal strategy for the auctioneer would entail submitting a bid of $z = \overline{x}(\tilde{\theta}_k,k)) + E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k))$ in each time period. By the same reasoning as in the proof of Theorem 8, it follows that the difference between the auctioneer's expected payoff from making a bid of $\overline{x}(\tilde{\theta}_k,k))$ and the auctioneer's expected payoff from making a bid of $z = \overline{x}(\tilde{\theta}_k,k)) + E_{\tilde{\theta}_{k+1}}[V_{k+1}(\overline{x}(\tilde{\theta}_{k+1},k+1))]-V_k(\overline{x}(\tilde{\theta}_k,k))$ is $\frac{f(\overline{x}(\tilde{\theta}_k,k))\delta^2(\Delta V)^2}{2(1-\delta)}+o((\Delta V)^2)$. Since the auctioneer's payoff from using the algorithm that we have outlined is also $\frac{f(\overline{x}(\tilde{\theta}_k,k))\delta^2(\Delta V)^2}{2(1-\delta)}+o((\Delta V)^2)$, it then follows that the difference between the auctioneer's payoff under the algorithm we have outlined and the maximum possible payoff the auctioneer could obtain under the theoretically optimal strategy is $o((\Delta V)^2) = o(\frac{1}{k^4})$.

But we know from Theorem 8 that the difference between the auctioneer's payoff under the algorithm we have outlined and the auctioneer's payoff under the greedy strategy is $\frac{\delta^2}{8(1-\delta)^3k^4}s^4(\overline{x}(\tilde{\theta}_k,k))f^2(\overline{x}(\tilde{\theta}_k,k)) + o(\frac{1}{k^4})$. From this it follows that the difference between the auctioneer's payoff under this strategy and the maximum possible payoff the auctioneer could obtain under the theoretically optimal strategy becomes vanishingly small compared to the difference between the auctioneer's payoff under this strategy and the auctioneer's payoff under the greedy strategy for large k. □

The results in the previous theorems suggest that the maximum possible payoff increase that can be achieved by incorporating active exploration into a machine learning system for online auctions is quite small for auctions involving ads that have already received a large number of impressions. However, in many auctions it is frequently the case that there are advertisers that have only received a small number of impressions, so it is desirable to know whether these conclusions for ads that have received large numbers of impressions will also hold for ads that have only received a small number of impressions. We present a result addressing this question next in Theorem 10.

THEOREM 10. *Suppose the bidder with unknown eCPM has a CPC bid of* 1 *and a click-through rate drawn from*

a beta distribution with parameters α and β. Also suppose that this bidder's expected eCPM is ω and the standard deviation in this bidder's true eCPM is $\gamma\omega$. Then the difference between the maximum possible payoff the auctioneer could obtain under the theoretically optimal strategy and the auctioneer's payoff from the greedy strategy is no greater than $\frac{\delta^2\gamma^8\omega^6\overline{f}^3}{8(1-\delta)^3(1-\omega)^2}$, where $\overline{f}$ denotes the supermum of $f(\cdot)$.

The proof of Theorem 10 is lengthy and relegated to the appendix of the full version of the paper [23]. Theorem 10 presents bounds on the maximum performance improvement that can be achieved over the purely greedy strategy by using active learning, but it is not immediately clear from this result whether these bounds imply there are significant limitations on the performance improvement that can be achieved by using active learning. We thus seek to shed some light on this under empirically realistic values of the parameters.

If the typical eCPM bids for the winning advertisers in an auction are roughly $\xi\omega$, then the auctioneer's total payoff for the game will be roughly $\frac{\xi\omega}{1-\delta}$, and the result in Theorem 10 indicates that the maximum percentage increase in expected payoff that one can achieve as a result of using the theoretically optimal strategy rather than the greedy strategy is on the order of $100\%\frac{\delta^2\gamma^8\omega^5\overline{f}^3}{8\xi(1-\delta)^2(1-\omega)^2}$.

Furthermore, if the typical eCPM bids for the highest competing advertisers in an auction are roughly $\xi\omega$, then $\overline{f}$ is likely to also be on the order of $\frac{1}{\xi\omega}$. This holds, for example, if the highest competing eCPM bids are drawn from a lognormal distribution, as the largest value of the density of a lognormal distribution with parameters μ and σ^2 is equal to $\frac{c(\sigma^2)}{\xi\omega}$, where $\xi\omega$ is the expected value of the lognormal distribution and $c(\sigma^2) \equiv \frac{e^{\sigma^2}}{\sqrt{2\pi\sigma^2}}$ is a constant that depends only on σ^2. Furthermore $c(\sigma^2)$ is likely to be close to 1 for realistic values of σ^2 since $c(\sigma^2) \in [0.93, 1.09]$ for values of $\sigma^2 \in [0.2, 1]$. The lognormal distribution is a realistic representation of the distribution of highest competing bids in online auctions since both [27] and [33] have noted that the distribution of highest bids can be well-represented by a lognormal distribution using data from sponsored search auctions at Yahoo!.

By using the facts that the value of $\overline{f}$ is likely to be on the order of $\frac{1}{\xi\omega}$ and the maximum percentage increase in expected payoff that one can achieve as a result of using the theoretically optimal strategy rather than the greedy strategy is on the order of $100\%\frac{\delta^2\gamma^8\omega^5\overline{f}^3}{8\xi(1-\delta)^2(1-\omega)^2}$, it then follows that the maximum percentage increase in expected payoff that one can achieve as a result of using the theoretically optimal strategy rather than the greedy strategy is on the order of $100\%\frac{\delta^2\gamma^8\omega^2}{8\xi^4(1-\delta)^2(1-\omega)^2}$.

There is empirical evidence that indicates that the typical click-through rates for ads in online auctions tend to be on the order of $\frac{1}{100}$ or $\frac{1}{1000}$ [10], so $(1-\omega)^2$ will be very close to 1 and ω^2 is likely to be less than 10^{-4} (for search ads) or 10^{-6} (for display ads). Furthermore, even for a brand new ad, the typical errors in a machine learning system's predictions are unlikely to exceed 30% of the true click-through rate of the ad, so $\gamma \leq 0.3$ is likely to hold in most practical applications. Finally, ξ is a measure of by how much the highest bid in an auction exceeds the typical eCPM bid of an average ad in the auction. Since there are normally hundreds of ads competing in online auctions, it seems that one can conservatively estimate that $\xi \geq 3$ is likely to hold in most real-world online auctions.

By combining the estimates in the previous paragraph, it follows that $100\%\frac{\gamma^8\omega^2}{8\xi^4(1-\omega)^2}$ will almost certainly be less than $10^{-9}\%$ in search auctions and $10^{-11}\%$ in display auctions. Now if $\delta \leq 0.9999$, $\frac{\delta^2}{(1-\delta)^2}$ will be no greater than 10^8, and if $\delta \leq 0.99999$, $\frac{\delta^2}{(1-\delta)^2}$ will be no greater than 10^{10}. Thus even for values of δ that are exceedingly close to 1 ($\delta = 0.9999$ for search ads and $\delta = 0.99999$ for display ads), $100\%\frac{\gamma^8\omega^2}{8\xi^4(1-\omega)^2}\frac{\delta^2}{(1-\delta)^2}$ will be no greater than 0.1%. Thus as long as $\delta \leq 0.9999$ (or $\delta \leq 0.99999$ for display auctions), the bound given in Theorem 10 guarantees that under empirically realistic scenarios, the maximum possible performance improvement that can be achieved by incorporating active learning into a machine learning system is at most a few hundredths of a percentage point. This is a finite sample result that does not require a diverging number of impressions in order to hold.

6. SIMULATIONS

The results of the previous section suggest that the overall benefit that can be obtained by incorporating active exploration into a machine learning system in an auction environment is exceedingly small. We now seek to empirically verify that the benefit that can be obtained from active exploration is indeed quite small by conducting simulations under some empirically realistic scenarios.

To do this, we consider a scenario in which there is a repeated auction in which a cost-per-click (CPC) bidder competes against a CPM bidder in each auction. The CPC bidder has a CPC bid of 1 and a fixed unknown click-through rate for all periods that is a random draw from the beta distribution with parameters α_C and β_C. The CPM bidder's CPM bid varies from period to period, and in each period we assume that the CPM bidder's CPM bid is a random draw from the beta distribution with parameters α_M and β_M. We let $f(\cdot)$ denote the probability density function corresponding to this distribution. Throughout we assume that payoffs are discounted at a rate of $\delta = 0.9995$ and that there are $T = 10000$ time periods.

While the CPM bidder's bid is drawn from the same distribution in every period, the auctioneer's beliefs about the distribution from which the CPC bidder's click-through rate is drawn changes over time. In particular, just before the auction in period t, the auctioneer believes that the CPC bidder's true click-through rate is a random draw from the beta distribution with parameters $\alpha_{t,C}$ and $\beta_{t,C}$ where $\alpha_{t,C}$ is equal to α_C plus the number of clicks the CPC bidder has received so far and $\beta_{t,C}$ is equal to β_C plus the number of times the CPC bidder's ad was shown but did not receive a click.

We compare total social welfare under two possible scenarios. The first scenario we consider is a standard ranking algorithm in which the ads are ranked purely on the basis of their expected eCPM bids. The second scenario we consider is one in which the CPC bidder makes a bid of the form $\overline{x}_t + \frac{\delta(1-\delta^{T-t})}{2(1-\delta)}\frac{\alpha_{t,C}\beta_{t,C}}{(\alpha_{t,C}+\beta_C)^2(\alpha_{t,C}+\beta_{t,C}+1)^2}f(\overline{x}_t)$ for the CPC bidder in each period t, where $\overline{x}_t$ denotes the CPC bidder's expected click-through rate just before the auction in period t, and the CPM bidder bids in the same way as

in the first scenario. This second scenario corresponds to adding a term equal to the value of learning to the CPC bidder's expected eCPM bid in the game with finite time horizons.

Throughout we focus on scenarios that are motivated by empirical evidence on the likely expected click-through rates for ads in online auctions. In particular, since empirical evidence indicates that the typical click-through rates for ads in online auctions tend to be on the order of $\frac{1}{100}$ or $\frac{1}{1000}$ [10], we focus on situations in which the expected click-through rate of the CPC bidder is small. Thus in all the simulations we conduct, we assume that the CPC bidder's expected click-through rate is on the order of $\frac{1}{100}$. We further assume that the CPM bidder's expected CPM bid is also on the order of $\frac{1}{100}$ in each auction.

Similarly, since it is unlikely that there will be substantial errors in the estimate of a new ad's predicted click-through rate, we also focus on situations in which there is only moderate uncertainty in the click-through rate of a new ad. In particular, throughout we consider distributions of the CPC bidder's bid such that the standard deviation in the advertiser's click-through rate is no greater than 20 or 30% of the expected value. However, since there is likely to be considerable variation in the distribution of competing CPM bids that an advertiser faces, we focus on distributions of the CPM bidder's bid for which the variance in this bid is quite substantial.

Conditions	Efficiency Increase
$\alpha_C = 10, \beta_C = 1000, \alpha_M = 2, \beta_M = 100$	-0.0011% (0.0149%)
$\alpha_C = 10, \beta_C = 1000, \alpha_M = 2, \beta_M = 200$	-0.010% (0.032%)
$\alpha_C = 20, \beta_C = 2000, \alpha_M = 2, \beta_M = 100$	0.0086% (0.0087%)
$\alpha_C = 20, \beta_C = 2000, \alpha_M = 2, \beta_M = 200$	0.013% (0.017%)

Table 1: Average percentage increase in efficiency from incorporating active learning into a machine learning system (with standard errors in parentheses) after 2500 simulations. None of these results are statistically significant at the $p < .05$ level.

Table 1 reports the results of the simulations that we have conducted. The conclusions from these simulations are striking. While we have conducted enough simulations to estimate the efficiency gain that can be obtained from adding active exploration to within a few hundredths of a percentage point, none of the resulting estimated efficiency gains realized in Table 1 are statistically significant. Indeed one can conclude from all of these simulations that the maximum possible efficiency gain that could be achieved in these settings is at most a few hundredths of a percentage point. These empirical results provide further support for our theoretical conclusions that the value of adding active exploration to a machine learning system in an auction setting is exceedingly small.

The reason for the results observed in Table 1 is that an optimal exploration algorithm in these auction settings will only do a tiny additional amount of exploration compared to a purely greedy strategy of simply always submitting a bid for the CPC bidder equal to the CPC bidder's bid. For instance, for the first simulation considered in Table 1, the incremental increase in an advertiser's bid in the first period of the game as a result of active exploration is only 3.6%, implying only about a 1.3% increase in the probability that the CPC bidder will be shown as well as only about a 1.8% increase in expected payoff conditional on the auctioneer showing a different ad under active exploration than under the purely greedy strategy. Thus the incremental expected payoff increase that can be achieved by incorporating active exploration into an existing machine learning system in this auction setting is at most a few hundredths of a percentage point.

The results in Table 1 make use of distributions that we regard as empirically realistic in the sense that there is a realistic amount of uncertainty about the click-through rate of the CPC bidder as well as a realistic amount of variation in the distribution of competing CPM bids. It is worth noting that if one relaxes the requirement that there be a realistic amount of uncertainty about these variances, then it is possible for the algorithm we have proposed to substantially outperform the purely greedy strategy of simply making a bid for the CPC bidder that always equals the CPC bidder's expected eCPM. In particular, if we instead assume that there is substantially more uncertainty about the CPC bidder's bid than we have assumed in the simulations in Table 1 and we also assume that there is substantially less variance in the distribution of competing CPM bids than we have allowed for in Table 1, then there will be considerably greater benefits to adding active exploration because there is both more to learn about the CPC bidder's true eCPM bid as well as less exploration that will take place for free solely due to random variation in the competing bids. In this case, there may well be significant benefits to adding active exploration to a machine learning system.

Conditions	Efficiency Increase
$\alpha_C = 2, \beta_C = 200, \alpha_M = 15, \beta_M = 1000$	0.70% (0.13%)
$\alpha_C = 5, \beta_C = 500, \alpha_M = 15, \beta_M = 1000$	0.20% (0.06%)

Table 2: Average percentage increase in efficiency from incorporating active learning into a machine learning system (with standard errors in parentheses) after 5000 simulations. These results are all statistically significant at the $p < .001$ level.

Table 2 reports the results of simulations that were conducted using distributions in which there is substantially more uncertainty about the CPC bidder's click-through rate and substantially less variance in the CPM bidder's competing CPM bid than in the distributions considered in Table 1. These simulations indeed reveal statistically significant efficiency gains as a result of active exploration. Nonetheless it is worth noting that the efficiency gains reported in Table 2 are still fairly small. Even when we make assumptions that bias the case in favor of active exploration being important, none of the efficiency gains reported in Table 2 are greater than a few tenths of a percentage point.

7. CONCLUSIONS

In online auctions there may be value to exploring ads with uncertain eCPM's to learn about the true eCPM of the ad and be able to make better ranking decisions in the future. But the online auction setting is very different from standard multi-armed bandit problems in the sense that there may be tremendous variation in the quality of competition that an advertiser with unknown eCPM faces in an auction, and as a result there will typically be plenty of free opportunities to explore an ad with uncertain eCPM in auctions where there simply are no ads with eCPM bids that are known to be high.

We have presented a model of the explore/exploit problem in online auctions that explicitly considers this random variation in competing bids that is present in real auctions. We find that the optimal solution for ranking the ads in this setting is dramatically different than the optimal solution in standard multi-armed bandit problems, and in particular, that the optimal amount of active exploration that results is considerably smaller than in standard multi-armed bandit problems. This in turn implies that the improvement in the auctioneer's expected payoff that can be achieved by adding active learning to a machine learning system in online auctions is also exceedingly small. Thus while it is theoretically possible to improve performance by incorporating active learning into a machine learning system for online auctions, in a practical exchange environment, a purely greedy strategy of simply ranking the ads by their expected eCPM's is likely to lead to nearly as strong a performance as any other conceivable strategy.

The model we have used in this paper considers a simple situation in which there a single advertiser with unknown eCPM that competes in each period against an advertiser with known eCPM whose eCPM bid is a random draw from some distribution. But our conclusions about the value of learning are not restricted to this simple model. In the full version of the paper [23] we present a variety of more complicated models including models in which there are multiple advertisers with unknown eCPMs who need to be ranked as well as models in which there is correlation between the unknown eCPMs of multiple different advertisers and information from showing one advertiser can help one refine one's estimate of the eCPM for some other advertiser. The substantive conclusions in this paper about the optimal bidding strategies and the value of learning all extend to a these more complicated models, so we expect our conclusion that about the value of active learning being quite small in a practical exchange environment to be robust to a variety of natural extensions of the model.

8. ACKNOWLEDGMENTS

We thank Joshua Dillon, Pierre Grinspan, Chris Harris, Tim Lipus, Mohammad Mahdian, Hal Varian, and Martin Zinkevich for helpful comments and discussions.

9. REFERENCES

[1] D. Agarwal, B.-C. Chen, and P. Elango. Explore/exploit schemes for web content optimization. In *Proceedings of the 9th Industrial Conference on Data Mining (ICDM)*, pages 1–10, 2009.

[2] P. Aghion, P. Bolton, C. Harris, and B. Jullien. Optimal learning by experimentation. *Review of Economic Studies*, 58(4):621–654, 1991.

[3] P. Aghion, M. P. Espinosa, and B. Jullien. Dynamic duopoly with learning through market experimentation. *Economic Theory*, 3(3):517–539, 1993.

[4] N. Anthonisen. On learning to cooperate. *Journal of Economic Theory*, 107(2):253–287, 2002.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[7] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC)*, pages 79–88, 2009.

[8] A. Banerjee and D. Fudenberg. Word-of-mouth learning. *Games and Economic Behavior*, 46(1):1–22, 2004.

[9] J. S. Banks and R. K. Sundaram. Denumerable-armed bandits. *Econometrica*, 60(5):1071–1096, 1992.

[10] E. Bax, A. Kuratti, P. McAfee, and J. Romero. Comparing predicted prices in auctions for online advertising. *International Journal of Industrial Organization*, 30(1):80–88, 2011.

[11] D. Bergemann and J. Välimkäki. Learning and strategic pricing. *Econometrica*, 64(5):1125–1149, 1996.

[12] D. Bergemann and J. Välimkäki. Market diffusion with two-sided learning. *RAND Journal of Economics*, 28(4):773–795, 1997.

[13] D. Bergemann and J. Välimkäki. Experimentation in markets. *Review of Economic Studies*, 67(2):213–234, 2000.

[14] D. Bergemann and J. Välimkäki. Stationary multi-choice bandit problems. *Journal of Economic Dynamics and Control*, 25(1):1585–1594, 2001.

[15] P. Bolton and C. Harris. Strategic experimentation. *Econometrica*, 67(2):349–374, 1999.

[16] M. Brezzi and T. L. Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108, 2002.

[17] S. Callander. Searching for good policies. *American Political Science Review*, 105(4):643–662, 2011.

[18] N. R. Devanur and S. M. Kakade. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC)*, pages 99–106, 2009.

[19] A. Fishman and R. Rob. Experimentation and competition. *Journal of Economic Theory*, 78(2):299–320, 1998.

[20] D. Gale. What have we learned from social learning? *European Economic Review*, 40(3-5):617–628, 1996.

[21] D. Gale and R. W. Rosenthal. Experimentation, imitation, and stochastic stability. *Journal of Economic Theory*, 84(1):1–40, 1999.

[22] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177, 1979.

[23] P. Hummel and R. P. McAfee. Machine learning in an

auction environment. *Google Inc. Typescript*, 2013.

[24] G. Keller and S. Rady. Optimal experimentation in a changing environment. *Review of Economic Studies*, 66(3):475–503, 1999.

[25] G. Keller and S. Rady. Strategic experimentation with poisson bandits. *Theoretical Economics*, 5(2):275–311, 2010.

[26] G. Keller, S. Rady, and M. Cripps. Strategic experimentation with experimental bandits. *Econometrica*, 73(1):39–68, 2005.

[27] S. Lahaie and R. P. McAfee. Efficient ranking in sponsored search. In *Proceedings of the 7th International Workshop on Internet and Network Economics (WINE)*, pages 254–265, 2011.

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

[29] R. A. Lewis. Where's the 'wear-out?' online display ads and the impact of frequency. *Massachusetts Institute of Technology Typescript*, 2010.

[30] S.-M. Li, M. Mahdian, and R. P. McAfee. Value of learning in sponsored search auctions. In *Proceedings of the 6th International Workshop on Internet and Network Economics (WINE)*, pages 294–305, 2010.

[31] L. J. Mirman, L. Samuelson, and A. Urbano. Monopoly experimentation. *International Economic Review*, 34(3):549–563, 1993.

[32] G. Moscarini and L. Smith. The optimal level of experimentation. *Econometrica*, 69(6):1629–1644, 2001.

[33] M. Ostrovsky and M. Schwarz. Reserve prices in internet advertising auctions: A field experiment. *Stanford University Typescript*, 2009.

[34] M. Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

[35] A. Rusitchini and A. Wolinsky. Learning about variable demand in the long run. *Journal of Economic Dynamics and Control*, 19(5-7):1283–1292, 1995.

[36] K. H. Schlag. Why imitate, and if so how? a boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78(1):130–156, 1998.

[37] B. Strulovici. Learning while voting: Determinant of collective experimentation. *Econometrica*, 78(3):933–971, 2010.

[38] X. Vives. Learning from others: A welfare analysis. *Games and Economic Behavior*, 20(2):177–200, 1997.

[39] M. L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979.

[40] J. Wortman, Y. Vorobeychik, L. Li, and J. Langford. Maintaining equilibria during exploration in sponsored search auctions. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE)*, pages 119–130, 2007.